Rodrigo Soares Chaves

Agrupamento de documentos eletrônicos por meio de Sintagmas Nominais

Belo Horizonte
2013

Rodrigo Soares Chaves

Agrupamento de documentos eletrônicos por meio de Sintagmas Nominais

Projeto de dissertação apresentado ao Programa de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento da Universidade Fundação Mineira de Educação e Cultura, como requisito parcial à obtenção do título de Mestre em Sistemas da Informação e Gestão do Conhecimento.

Linhas de Pesquisa:
Sistemas da Informação

Orientador:
Prof. Dr. Luiz Claudio Gomes Maia

Belo Horizonte
2013

RESUMO

CHAVES, Rodrigo Soares. **Agrupamento de documentos eletrônicos por meio de Sintagmas Nominais**. 2013. Projeto de Dissertação (Mestrado Profissional em Sistemas da Informação e Gestão do Conhecimento) – Universidade Fundação Mineira de Educação e Cultura. Belo Horizonte, 2013.

Com o desenvolvimento tecnológico a informação passou a ter um papel fundamental em nossas vidas. A forma como a informação é recuperada, tratada e representada passa a ter enorme importância, assim como o tempo necessário para obtê-la. A classificação das informações é, na maioria das vezes, feita por processos manuais que exigem muito esforço por parte de bibliotecários e pesquisadores. A proposta deste trabalho é descrever as atividades experimentais realizadas com base nos sintagmas nominais para a análise, classificação e agrupamento de documentos eletrônicos de forma automática. Por meio de técnicas de agrupamento adotadas em similaridades de documentos, a pesquisa verifica os benefícios alcançados com a distribuição dos trabalhos de cada grupo. Como parâmetro para a realização das buscas é proposto a utilização de um corpus formado por aproximadamente 12.000 resumos de estudos acadêmicos, que serão submetidos a experimentos e análise.

Palavras-chave: agrupamento automático de documentos, similaridade de documentos, análise de texto, sintagmas nominais.

SUMÁRIO

_

1. INTRODUÇÃO

O uso da informação é cada vez mais comum nos meios digitais. A criação e a divulgação da informação tornaram-se atividades triviais, o que provocou um grande volume de textos publicados e acentuou a importância das coletas eficazes de informação. Por mais que novas tecnologias sejam desenvolvidas, informações irrelevantes ou de baixa qualidade continuam a ser encontradas, abrindo espaço para um campo de pesquisa responsável pela recuperação da informação (MAIA, 2008; DIAS, 2001).

Sistemas de Informação, de forma geral, viabilizam novas atividades sociais e profissionais que já estão inseridas no nosso cotidiano em inúmeras atividades humanas (SOUZA; ALVARENGA, 2004). Quando se trata de Sistemas de Recuperação da Informação (SRI), o objetivo é a realização de buscas de informações relevantes em um determinado contexto a partir de determinados parâmetros passados no momento da pesquisa. Com os rápidos processamentos realizados pelos computadores de hoje, algoritmos destinados a esta recuperação possibilitaram o surgimento de novas técnicas e consequentemente uma considerável melhora na qualidade e relevância dos resultados obtidos (MAIA, 2008).

A maior parte dos SRIs atualmente disponíveis utiliza a palavra como unidade básica de consulta à informação (KURAMOTO, 1996; WIVES, 1999; ALVARENGA, 2001; SOUZA; ALVARENGA, 2004; MAIA, 2008). Essa utilização de palavras pode causar interpretações equivocadas por parte do SRI quanto ao real significado expressado pelo usuário no momento de solicitar informações acerca de um tema. Na maioria das consultas a documentos eletrônicos também são utilizadas palavras-chaves, uma ou duas, para um universo de informações (NIELSEN; LORANGER, 2007).

Os SRIs construídos em ambientes distribuídos, como a internet, passam por dificuldades na identificação de informações relevante aos usuários, devido entre outros fatores, ao grande volume de informações contidas na rede. A internet, embora construída com características anárquicas, apresenta um imenso repositório de documentos eletrônicos, que possibilita acesso rápido à informação desejada por meio de mecanismos de busca (SOUZA; ALVARENGA, 2004). Entretanto, os dados na web são de difícil compreensão e estruturação, o que pode

provocar uma queda da qualidade e precisão das buscas (ANTONIOU, 2004). Nem sempre o resultado retornado é o esperado e o desejado pelo usuário, principalmente se os documentos estiverem contidos em arquivos eletrônicos e não em páginas web.

A maioria dos SRIs que operam atualmente na internet, recebem um termo como argumento de pesquisa para recuperar informações contidas em páginas web, o que exige, por parte dos usuários, um conhecimento prévio acerca do que se espera com a consulta. O motivo é o fato de que os resultados estão diretamente relacionados com os termos utilizados na montagem do critério das buscas. Estes termos, são, na maioria das vezes, formados por referências para objetos existentes no mundo real (KURAMOTO, 2002). Os Sintagmas Nominais (SN) são apresentados neste contexto como unidade básica de representação da informação, sendo uma alternativa para o tratamento da informação. Pela análise semântica com base nos sintagmas nominais é possível obter novos parâmetros de buscas. Em alguns casos, os métodos baseados em Sintagmas Nominais apresentam maior eficácia se comparadas às análises baseadas em palavras-chave. Assim como as palavras e termos, os sintagmas nominais podem ainda ser usadas como descritores em outros processos de busca (SOUZA; ALVARENGA, 2004).

Por sua vez, documentos eletrônicos podem apresentar diversos termos relacionados ao conteúdo descrito. A extração e a ponderação das palavras ou dos sintagmas nominais associados a um documento são atividades que até pouco tempo eram executadas manualmente. Com a tecnologia, análises automatizadas a documentos eletrônicos tornaram-se viáveis possíveis. Souza e Alvarenga (2004) propõem uma metodologia para escolha semiautomática de descritores para representar documentos textuais digitais, escritos no idioma português, com base nas estruturas textuais e semânticas conhecidas como sintagmas nominais. O autor destaca as chances que o método possui de se tornar seguro para a atribuição de descritores contidos em um documento.

Em seu trabalho, Maia (2008) verificou os aprimoramentos dos cálculos de medidas de similaridade entre documentos eletrônicos por meio de técnicas de processamento de linguagens naturais propostas por estudos anteriores envolvendo a ciência da informação, a computação e a linguística.

A grande dificuldade no processo de recepção e tratamento do conhecimento está associada à complexidade no ato de classificar as informações, independente de intervenção humana (ALVARENGA, 2001). Existe todo um trabalho intelectual por parte do autor em combinar as palavras de um texto atribuindo a elas significados específicos. Quando se utiliza as palavras como unidade básica de acesso à informação, estas voltam a ter significados genéricos, sem qualquer referência a um objeto ou fato da realidade, ao qual o autor se referenciou (KURAMOTO, 2002). As novas tecnologias possibilitam o uso de sistemas automatizados de recuperação da informação, além de envolver mudanças na forma de trabalho dos autores e bibliotecários quanto ao processo de produção, armazenagem, tratamento e recuperação de documentos e informações (ALVARENGA, 2001).

A utilização da semântica embutida nos documentos é pouco explorada pelos SRIs e as estratégias de busca estão fortemente atreladas ao idioma em questão (SOUZA; ALVARENGA, 2004), no caso o Português. A automatização de processos relacionados à análise e à comparação de documentos é tida como um campo pouco explorado no idioma português. Com o avanço tecnológico novas ferramentas são desenvolvidas voltadas para a análise textual, embora poucas trabalhem com SN e poucos apresentam interface disponível na web.

Este projeto consiste na construção e na realização de experimentos na ferramenta OGMA Web, uma adaptação da ferramenta OGMA, desenvolvida por Luiz Claudio Gomes Maia em 2008. Como principal objetivo o trabalho se restringe à solucionar a questão: De que maneiras o agrupamento de documentos eletrônicos utilizando os sintagmas nominais pode ser aplicado de forma eficiente em um conjunto de documentos?

Como **objetivo principal** é proposto a investigação e a utilização de sintagmas nominais como meios de agrupamento por similaridade de documentos eletrônicos. Para a satisfazer o objetivo principal acima mencionado, os objetivos específicos abaixo listados, fizeram-se necessários:

a) Realizar a conversão da ferramenta OGMA para a plataforma web;
b) Aprimorar recursos envolvidos com o agrupamento de documentos disponíveis na ferramenta OGMA.
c) Analisar o resultado do agrupamento de documentos envolvendo medidas de similaridade em determinado corpus.

Desafios constantes vêm sendo colocados aos pesquisadores no que diz respeito à criação de novos processos que possam ser compatíveis com a agilidade, capacidade de armazenagem e processamento de informações das maquinas (ALVARENGA, 2001). Com o surgimento de tecnologias voltadas para a informação e a comunicação, o volume de informação e o número de usuários cresceram rapidamente na internet (KURAMOTO,2002). Nota-se uma crescente preocupação com as formas de tratamento e organização da informação (MAIA, 2010). Em geral a maior parte do conteúdo gerado é esquecido por não passar pelo processo humano de leitura, entendimento e síntese do volume informacional. A ciência deve incentivar cada vez mais a criação e desenvolvimento de mecanismos de buscas por informações precisas e relevantes, que auxiliem as pessoas a potencializarem o conhecimento (MAIA, 2008).

Um dos objetivos que deve ser almejado por qualquer sistema de recuperação da informação é a proposição de métodos que permitam uma seleção eficaz e eficiente à informação necessária. Como desafio, os sistemas de recuperação da informação devem coletar, representar, organizar e recuperar documentos considerando outras formas de representação da informação (MAIA, 2008). Os bibliotecários ou gestores de acervos de texto vêm criando diferentes tipos de representações e atribuindo informações primárias a registros específicos relativos a esta informação. Esses conjuntos de informações, podem ser considerados metainformações ou metadocumentos (ALVARENGA, 2001). Por meio de processo computacionais de catalogação de acervos textuais eletrônicos com base no agrupamento por similaridade dos documentos eletrônicos é possível atingir um patamar tecnológico que viabilize de forma automatizada desde a análise do texto, extração e classificação dos sintagmas nominais até a categorização dos documentos.

Este trabalho tem como motivação a criação de uma ferramenta acessível via internet, que possa facilitar o acesso à informação contida em documentos digitais por meio de agrupamentos baseados na similaridade entre os mesmos.

2. REVISÃO SISTEMÁTICA DA LITERATURA

A qualidade de uma pesquisa acadêmica está diretamente relacionada aos trabalhos que fazem parte da sustentação teórica de um determinado assunto. A Revisão Sistemática da Literatura (RSL) neste contexto, é apresentada como forma de analisar e filtrar os estudos relevantes a um determinado tema (PETERSEN, 2008). Diferentemente da forma aleatória de seleção da literatura, a RSL consiste em uma análise crítica com base em critérios de seleção/inclusão dos trabalhos.

Este capítulo tem como objetivo identificar e classificar estudos relevantes ao tema "Uso de SN para Identificação e Classificação Automática de Documentos Eletrônicos" por meio de pesquisas criteriosas em SRIs que trabalham com mecanismos de buscas por palavras chave e por documentos eletrônicos.

Como se trata de um estudo linguístico, os idiomas abordados nos trabalhos encontrados foram levados em consideração na classificação dos documentos, assim como as características adotadas na elaboração dos textos e os termos mais relevantes extraídos por um processo automático de classificação de documentos, conforme proposta de Souza e Alvarenga (2004) e implementação de Maia (2008).

Por meio da escolha automática de palavras-chaves que representem um documento é possível submeter os termos mais relevantes como parâmetros de busca em outros mecanismos de pesquisa. Apresenta, assim, traços da Web Semântica (WS) que preza o cruzamento de dados entre de diferentes bases.

2.1 Estratégias de Busca

A ferramenta OGMA Web, implementada no decorrer deste estudo, faz uma relação dos SNs e da WS e fornece uma interface web de fácil acesso aos usuários. Com a combinação desses recursos é possível estabelecer uma contextualização dos trabalhos relacionados a um determinado documento eletrônico, que foi utilizado como insumo para o OGMA Web, com o objetivo de identificar possíveis desdobramentos do tema extraído pela ferramenta.

O documento adotado como ponto inicial da pesquisa foi: "*Uso de sintagmas nominais na classificação automática de documentos eletrônicos - LC Maia, RR Souza 2010*". A saída do programa OGMA Web apresenta uma lista de sugestões e, por meio destas sugestões propostas, foi estabelecida uma visualização gráfica dos trabalhos relacionados ao tema obtido pela ferramenta. Como o OGMA Web apresenta limitações quanto ao número de caracteres analisados em um documento, foi selecionado apenas o capítulo referente às considerações finais de cada artigo.

Os trabalhos sugeridos pela ferramenta foram então coletados, quando disponíveis, e lapidados para submissão a nova análise do OGMA Web de forma recursiva. Os trabalhos sugeridos pela análise foram, então, coletados e lapidados para uma nova submissão.

2.2 Critérios de Seleção

Como as pesquisas por documentos eletrônicos podem sugerir outros estudos que não se encontram disponíveis para leitura, a disponibilidade dos arquivos eletrônicos foi critério crucial para continuidade do mapeamento dos desdobramentos das abordagens dos trabalhos.

Segundo Castells (2003), a Internet surgiu nos anos 90 e se disseminou em 1995, quando alcançou 16 milhões de usuários. Apesar de não ser utilizado como critério de seleção para análise, os trabalhos com o ano de publicação posterior ou igual a 1995 obtiveram destaque.

2.3 Classificação

A classificação dos trabalhos sugeridos pelo OGMA Web baseou-se na ponderação dos termos encontrados pela ferramenta para uma ordenação dos SN mais relevantes ao texto. Como resultado da busca a partir de um documento eletrônico de origem, é esperado uma classificação que viabilize um agrupamento e um mapeamento dos estudos. Um grafo representa em um plano o relacionamento dos trabalhos conforme as sugestões do aplicativo web integrado com o mecanismo de pesquisa. Neste grafo, os documentos são representados pelos nós e as sugestões pelas associações ou arestas.

2.4 Resultados

Foram levantados 79 artigos por meio da análise automatizada de 16 documentos eletrônicos, assim como das referências pertinentes à estruturação dos títulos conforme as sugestões da

ferramenta. Os artigos foram dispostos na ferramenta Gephi, software de visualização e manipulação de dados, de forma a representar como nós os trabalhos e as sugestões como arestas.

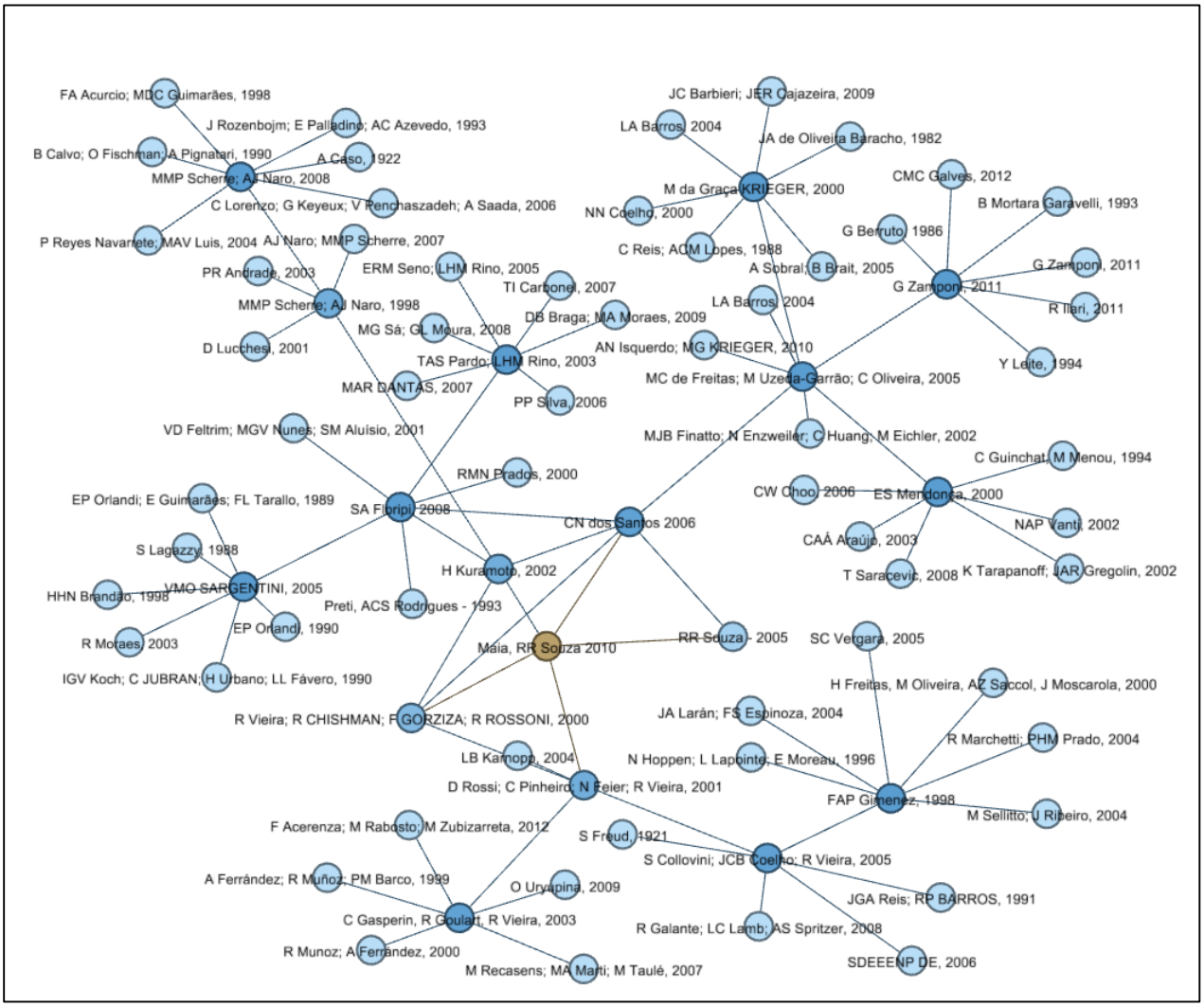

FIGURA 1- Rede de Grau
Fonte: Gephi 0.8.2 beta.

Os nós vistos em uma tonalidade mais escura, apresentam grau de relacionamento elevado se comparados aos nós na tonalidade mais clara. O estudo de Maia e Souza (2010), considerado aqui como ponto de partida para elaboração do grafo, está situado no centro. Como a proximidade entre um documento disposto no grafo e o documento de origem pode ser entendida como semelhança dos assuntos abordados pelos trabalhos, os estudos ligados diretamente ao nó de origem obtiveram destaque.

- *Uma abordagem alternativa para o tratamento ea recuperação de informação textual: os sintagmas nominais - H Kuramoto – 2002*;
- *Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro - CN dos Santos 2006;*
- *Resolução de correferência em textos da língua portuguesa - D Rossi, C Pinheiro, N Feier, R Vieira 2001.*

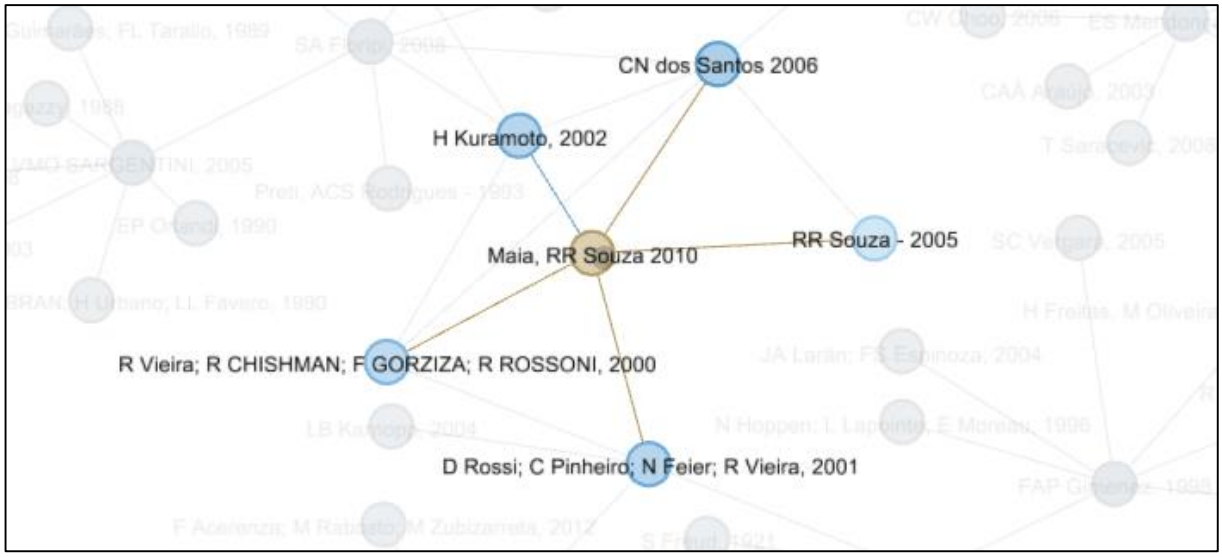

FIGURA 2 - Rede simples de trabalhos próximos à origem
Fonte: Elaborada pelo autor da dissertação.

Na FIGURA 2, são apresentados todas as relações primárias do documento de origem. O estudo *"Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais - RR Souza – 2005"*, não pode ser encontrado, já o estudo *"Extração de sintagmas nominais para o processamento de co-referência - R Vieira, R CHISHMAN, F GORZIZA, R ROSSONI 2000"*, não sugeriu novos trabalhos após o processamento da ferramenta.

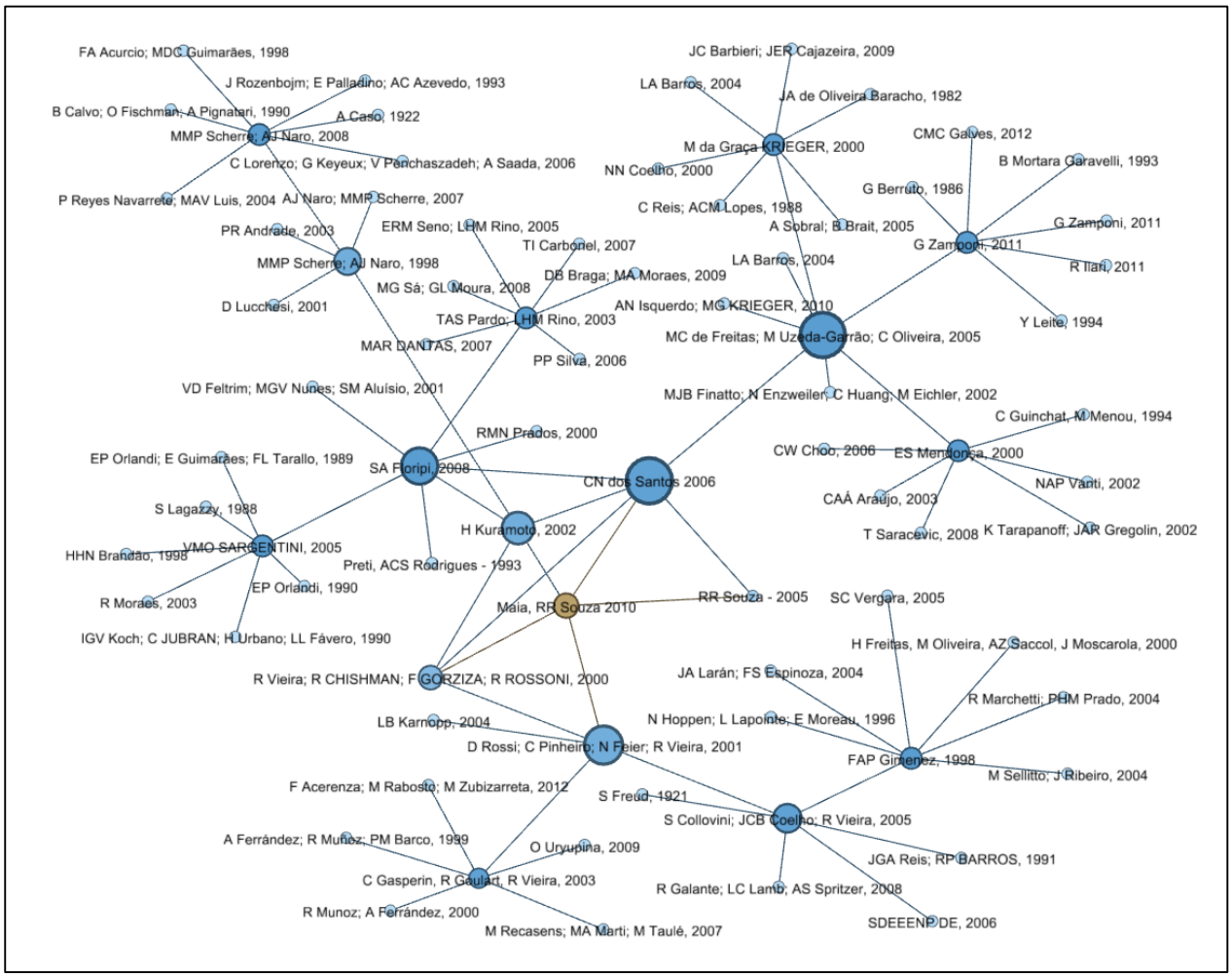

FIGURA 3 - Rede de proximidade
Fonte: Gephi 0.8.2 beta.

A análise da rede classificada por meio da proximidade entre os elementos do grafo foi feita com base no tamanho dos nós. Acentua a importância dos estudos já destacados com a análise dos graus dos nós e identifica os estudos "*Estudo da variação do determinante em sintagmas nominais possessivos na história do português - SA Floripi – 2008"* e *"Anotação de um corpus para o aprendizado supervisionado de um modelo de SN - MC de Freitas, M Uzeda-Garrão, C Oliveira, 2005;*" como relevante para o tema.

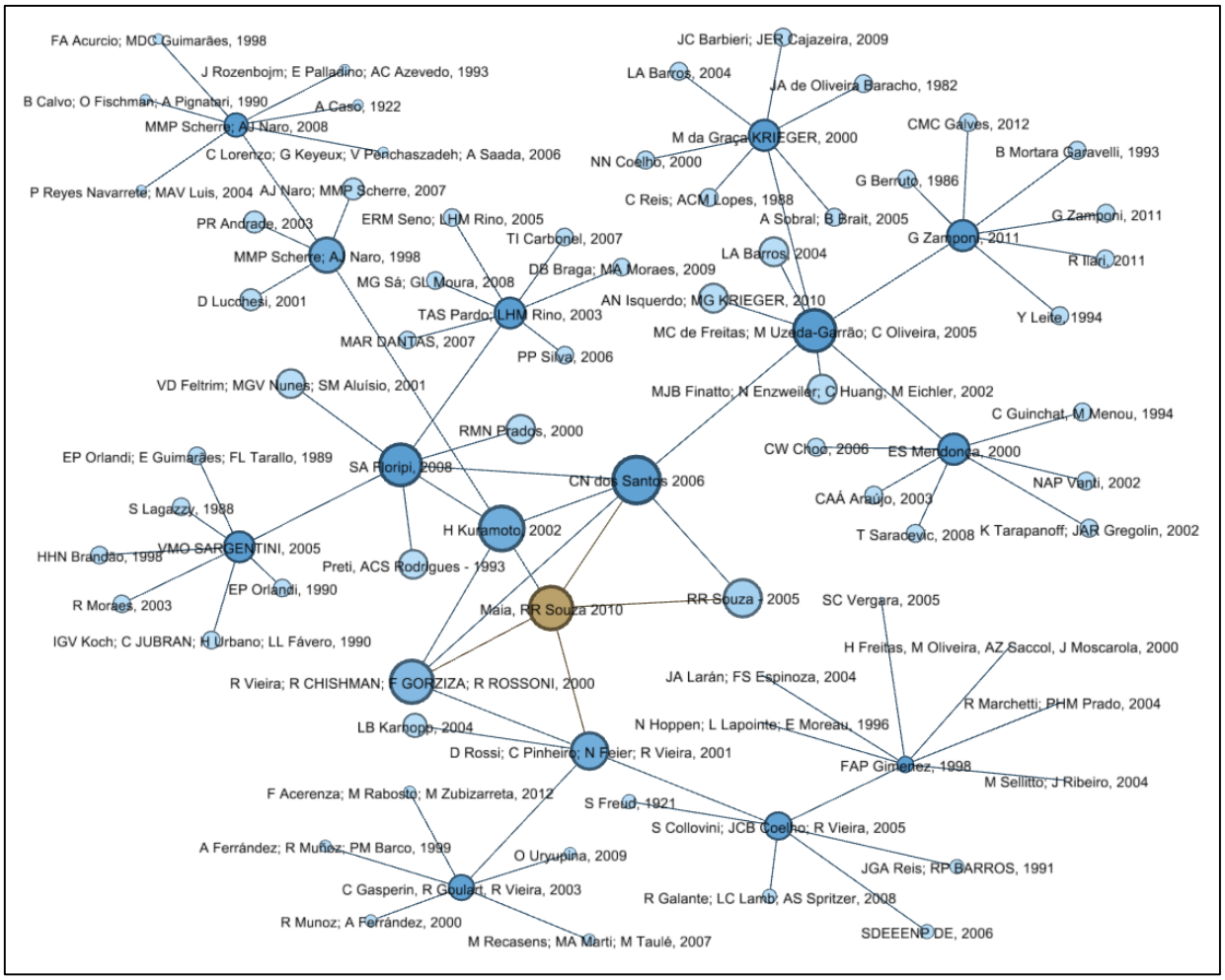

FIGURA 4 - Rede de intermediação
Fonte: Gephi 0.8.2 beta.

Com a disposição da rede conforme a intermediação, foi possível realçar os trabalhos próximos ao documento de origem, já listados anteriormente. Também foi possível observar a importância dos estudos: "*Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais - H Kuramoto - 2002"* e *"Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro - CN dos Santos 2006*" para o tema central da pesquisa.

Esta contextualização abordou o recurso de identificação de trabalhos relacionados a um documento eletrônico por meio de buscas realizadas de diferentes formas, pois foi utilizado tanto os SN quanto as palavras como unidade básica de informação. Com a integrações entre estes os mecanismos de busca, foi constatado que o tema central '*Sintagmas Nominais'* se dilui em outras áreas de estudos paralelos à medida que se distância do documento de origem. Foi possível identificar possíveis desdobramentos traçados a partir do artigo escrito por Maia

(2008), em diferentes áreas do conhecimento como: Processamento da Linguagem Natural; História Sociolinguística do Português no Brasil; Mineração de Dados; Análise de Redes Sociais; Análise Linguística de Idiomas Estrangeiros; Inteligência Organizacional e Gestão do Conhecimento.

3. REFERENCIAL TEÓRICO

3.1 Recuperação da Informação

O uso da informação é cada vez mais comum nos meios de comunicação digitais (MAIA, 2008). Com a globalização, a informação passou a ocupar cada vez mais importância para a sociedade de forma geral (DANTAS, 1987). Com isto, a criação e a divulgação da informação tornaram-se atividades triviais, o que provocou, em consequência, um grande volume de textos publicados e acentuou, ainda mais, a importância de coletas eficazes de informação (MAIA, 2008; DIAS, 2001). O sucesso do processo de recuperação da informação está associado diretamente ao tempo gasto pelo usuário para alcançar os resultados desejados (BASKERVILLE; DULIPOVICI, 2006). A internet exemplifica bem este cenário. Construída com características anárquicas, ela apresenta um imenso repositório de documentos eletrônicos, que possibilita o fácil acesso e a recuperação da informação por meio de mecanismos de busca (SOUZA; ALVARENGA, 2004). Porém, informações irrelevantes ou de baixa qualidade continuam a ser encontradas (MAIA, 2008) e, com isso, é descoberto um campo de pesquisa referente à recuperação da informação (DIAS, 2001).

A recuperação da informação, tem como um dos desafios, atender às necessidades de específicas do usuário de forma rápida e precisa. A maioria dos métodos atualmente utilizados na recuperação da informação colocam a palavra como unidade básica de acesso à informação (KURAMOTO, 2002). Apesar de apresentarem melhorias significativas nos resultados retornados, as buscas por palavras-chaves ainda mostram consideráveis limitações (ALVARENGA, 2001). Sistemas de recuperação de informação usualmente adotam termos índices formados por palavras-chave (SOUZA; ALVARENGA, 2004). Os descritores contidos em um documento são termos portadores de informações que fazem referência a objetos do mundo real. Prem, com o crescente volume de informações, tornou-se necessária a elaboração de mecanismos de indexação automática, que se baseiam na extração de palavras, inadequadamente tidas como descritores, mas que não garantem que a informação foi completamente extraída (KURAMOTO, 2002). Com as estratégias de representação e indexação de documentos, há uma vertente que busca levar em conta a semântica intrínseca textual, através de sintagmas nominais como descritores, no lugar das palavras-chave (SOUZA, 2004).

A informação é tida de forma interpretativa e, a princípio, não está formalizada em símbolos que representem algum significado. Eventualmente pode ocorrer, com a presença de dados, o que possibilita a utilização de um computador com o intuito de processamento, mas sem descartar o entendimento a qual a informação está sendo passada (SETZER, 1999). A informação consiste em fatos relacionados a dados que são organizados para descrever uma determinada situação ou condição (BASKERVILLE; DULIPOVICI, 2006).

A informação é um componente associada a quase todas as organizações no planeta, em diferentes situações. Para sua recuperação são necessárias a compreensão dos processos cognitivos que estão vinculados à informação, à percepção transformada pela informação e a importância das fontes tecnológicas para esta recuperação, que são fundamentais em muitos sistemas sociais e em atividades humanas (CHOO, 2003). A ciência da informação, perante a tecnologia, pode ter várias definições e classificações. Pode ser dividida em teoria da informação, processos de comunicação, representação da informação, teoria da classificação e armazenamento e recuperação da informação (DIAS, 2000). Com o objetivo de tornar acessível o conhecimento a quem necessita, a compreensão da recuperação da informação deve ser abordada juntamente com as atividades que a possibilitam, nas relações e nas ações que ocorrem durante seu processo (GARCIA, 2007).

Segundo Renato Souza e Lídia Alvarenga (2004), em 1962 foi proposta por Ted Nelson e Douglas Engelbart uma interface imaginária, em que textos poderiam estar associados a outros textos, tornando assim uma navegação entre textos feitas de forma intuitiva, criava-se assim o conceito de hiperlink. Com a internet surgindo nos anos 90, o estrutura textual pensada por Ted e Douglas veio a acontecer de forma abrangente pelo mundo, o que desencadeou uma explosão de documentos disponíveis na rede, e diversos sistemas de informações com o intuito de recuperação da informação. Porém, a internet se tornou um grande repositório de informações não catalogadas e de difícil localização (SOUZA; ALVARENGA, 2004).

Em 1978, Lancaster já anunciava a mudança na forma como deveriam ser tratadas as informações referentes a documentos, que passariam a ser tratadas de forma eletrônica, com o intuito de facilitar o trabalho de pesquisa e tornar mais simples o aprendizado.

Segundo Hélio Kuramoto (1995), os recursos informacionais estão se tornando mais acessíveis a cada ano. Porém, ainda existe uma dificuldade para acessar estes recursos de forma fácil e precisa. Para facilitar este acesso, são utilizados os Sistemas de Recuperação da Informação (SRIs). Para Lancaster e Warner (1993) os SRIs são formados por interfaces entre usuários e coleções de informações, considerando a recuperação, o armazenamento, os critérios de seleção, a análise do próprio usuário, o método de busca da informação e a automatização deste processo.

O mesmo foi constatado por Eduardo Wense Dias, em 2001, quando declarou que os meios tradicionais de troca de informação estão se tornando automatizáveis. Passamos a utilizar mais os documentos eletrônicos, mesmo que os papeis impressos ainda são comumente usados no dia a dia. Com o passar do tempo foram utilizados cada vez mais os sistemas de recuperação da informação, que a cada vez se tornavam mais complexos, e apresenta melhores precisões nas buscas.

Temas como sistemas de informação e de recuperação da informação vem sendo trabalhados mais a fundo por pesquisadores, a partir do século XX. Porém, existem diversas classificações de documentos técnicos, gerais e especializados do conhecimento, onde a classificação do documento se faz necessária com a análise humana, baseada no conhecimento preestabelecido do tema abordado, como por exemplo, física, química, direito, e mais recentemente computação (DIAS, 2001).

> Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a Web foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade (SOUZA; ALVARENGA, 2004, p 133).

A tecnologia é vista como um facilitador aos acessos às coleções de informações, embora que estas informações já existiam antes da internet. A web, neste contexto, apresenta imensa vantagem no tratamento e na recuperação da informação. A complexidade da análise textual executado por um SRI está diretamente ligado com o detalhamento desejado ou de acordo com o a metodologia de recuperação adotada (DIAS, 2001). Todo este processo representa um

enorme campo de pesquisa, que vem considerando cada vez mais características da língua e do vocabulário em busca de um processamento mais preciso e refinado.

Sob a perspectiva da ciência da informação, o uso intensivo das tecnologias provoca mudanças significativas na forma como o conhecimento é passado adiante, principalmente no que diz respeito a representação, a armazenagem e a recuperação das informações por mecanismos eletrônicos perante a cognição humana (ALVARENGA,2003).

Maia (2008), em seu trabalho, ressalta que: "Há muitos anos o homem tem armazenado, catalogado e organizado a informação, com o principal objetivo de recuperá-la para uso posterior ... ". Sistemas de recuperação da informação são criados para lidar com ciclos ininterruptos de criação e demanda de informação, por meio de tecnologias mecânicas e digitais de computação que são capazes de gerenciar grandes acervos de documentos (SOUZA; ALVARENGA, 2004).

Com a difusão das ferramentas de divulgação e armazenamento em meios digirais, tornou-se praticamente impossível a estruturação adequada das informações sem o auxílio de ferramentas digitais (MOREIRA; ALVARENGA, 2004). Com a evolução contínua, os sistemas informacionais deixaram de ser apenas uma interface para os usuários controlarem bancos de dados, e passaram a envolver não só a navegação, como também a forma como os dados são consultados e recuperados. Se combinadas, estas interações com a internet alcançamos novos padrões de bases de dados. Obtendo assim um patamar tecnológico avançado, capaz de mesclar informações para atingir um determinado resultado anteriormente impensável (ANTONIOU, 2004). Estes conceitos foram denominados como Web Semântica (WS), que possibilita o uso sofisticado da informação (BERNERS-LEE et al., 2001). Com a utilização da WS, novas formas de recuperar a informação são desenvolvidas e novos resultados são obtidos.

A melhoria da eficácia dos sistemas de recuperação está associada a diversas linhas de pesquisa como: a) a exploração das informações semânticas e semióticas intrínsecas aos dados; b) novas possibilidades de marcação semântica dos dados utilizando-se das metalinguagens; c) estratégias de apresentação da informação; d) perfis personalizados de utilização para determinado usuário (SOUZA, 2004).

3.2 Análise Textual

A expansão dos acervos digitais exigiu uma adaptação dos índices e pontos de acesso para garantir a recuperação bem sucedida da informação (ALVARENGA, 2001).

O tratamento de texto não é suficiente para garantir a recuperação eficaz da informação, a análise da linguagem humana feita pelo computador é fundamental para a recuperação bem sucedida de documentos, este procedimento é conhecido como Processamento da Linguagem Natural (PLN) (MAIA, 2010).

Para uma abordagem completa da organização e da recuperação da informação, por uma ótica textual, são adotadas as seguinte estratégias: a) indexação dos documentos mais significativos; b) forma adequada de apresentar as informações recuperadas aos usuários; c) utilização de padrões de metadados; d) adaptação contínua do sistema (SOUZA, 2004).

A representação pode ser interpretada como uma instância do processo cognitivo humano que é entendido por uma representação primária do conhecimento, como a matéria prima do registro mental, e do suporte documental. Envolve as etapas de percepção, identificação, interpretação, reflexão e codificação ao tomar conhecimento de algo novo, a partir dos sentidos, da emoção, da razão e da linguagem (ALVARENGA,2003). Vickery (1986) afirma que toda a representação de documento é simbólica. Por uma ótica computacional, as metas de trabalho não são apenas à criação de representações simbólicas dos documentos, apesar de englobar a criação de novas formas de escritas para hipertextos, resultando assim em novas representações (ALVARENGA,2003).

Alvarenga (2003) também apresenta os conceitos: Linguagem Especializada, que atua em um domínio restrito; Linguagem Normalizada, uma linguagem controlada; Unidades Linguísticas, que são os termos propriamente dito e Linguagem Pré e Pós Coordenada, que indicam se os termos são ou não combinados no momento do processamento. Para Maia (2010), a mineração de texto constitui na extração de informações sobre padrões em grandes bases de documentos textuais.

Tesauros podem ser considerados como ontologias simples, já que as ontologias dependem de uma complexidade nas relações apresentadas. Tem como função a representação dos assuntos dos documentos de modo a realizar análise a partir da indexação e da identificação do conteúdo, e em seguida "traduzir" para termos legíveis ao tesauro (MOREIRA; ALVARENGA, 2004).

Com o intuito de formar bases de conhecimento interoperáveis e bem estruturadas foi definido o termo ontologias (MOREIRA; ALVARENGA, 2004). Uma ontologia é uma especificação explícita e formal de uma conceituação (STUDER., 2001). Pode ser vista também como um modelo conceitual e abstrato em sua aplicação na informação (STAAB,2009).

Podemos afirmar então que para a elaboração e a construção de ontologias devemos levar em consideração a relação com a arquitetura do sistema de informação em que ela está envolvida com o objetivo de formular teorias de conhecimento para um domínio específico (JIMÉNEZ, 2004).

Os portais na internet também podem ser vistos em um contexto do uso das ontologias, uma vez que possuem uma estrutura multidimensionadas e apresentam além de um conteúdo midiático, imagens, documentos, ferramentas inteligentes, interfaces comerciais (JIMÉNEZ, 2004).

3.2.1 Processamento da Linguagem Natural

O tratamento de texto não é suficiente para garantir a recuperação eficaz da informação desejada. A análise feita pelo computador acerca da linguagem humana é fundamental para uma recuperação precisa de documentos eletrônicos, este procedimento é conhecido como processamento da linguagem natural (PLN) (MAIA, 2010).

A PLN tem como objetivo tratar os aspectos da comunicação humana por meio de processamentos automatizados realizados pelo computador. Faz com que o computador comunique com a linguagem humana. Porém, não é possível um entendimento de todos os níveis de comunicação, como sons, palavras, sentenças e discursos (MAIA, 2008).

A análise da semântica latente (LSA) pode ser incorporada à PLN. Manipulando os vetores de índice de um texto, empregando a matemática para relacionar os termos e decompor os vetores de índices. Maia (2008), exemplifica a LSA com uma suposta análise da frase: ‘extravio de bagagem’. A LSA que trabalha com a sinonímia e a polissemia neste caso irá considerar os termos: ‘extravio de bagagem’ e ‘extravio de mala’, uma vez que ‘bagagem’ e ‘mala’ possuem o mesmo significado no contexto. Outro exemplo é a frase: ‘banco de dados’, em que o sistema deve desconsiderar o entendimento da palavra ‘banco’ para entidades financeiras.

3.2.2 Índices de Pesos (TF - IDF)

Estes índices de pesos são atribuídos levando em consideração a frequência em que um termo é utilizado em um texto, o que possibilita a identificação da relevância de uma palavra, ou um sintagma em relação à um determinado documento. A relevância pode aumentar ou diminuir de acordo com o número de vezes que o termo é empregado no texto levando em consideração o número de termos totais utilizados no documento.

TF, ou *term frquency*, representa o número de vezes em que um termo aparece em um determinado documento. Para melhor representar este índice, é realizada uma normalização com base no número total de termos contidos no texto, o que garante uma correta ponderação do termo, independentemente do tamanho do documento.

$$TF = \frac{NTermos}{Total}$$

O cálculo da frequência é obtido com a divisão do número de vezes que o termo é empregado no documento (NTermos) pelo total de termos contidos no documento(Total).

Já o índice IDF, avalia a relevância de um termo em uma coleção de documentos. É calculada por meio da divisão do total de documentos da coleção (TotalDoc) pelo número de documentos que contém o termo analisado (DocTermos).

$$IDF = \frac{TotalDoc}{DocTermos}$$

3.2.3 Sintagmas Nominais

Um dos objetivos da recuperação da informação é estudar e propor métodos que permitem uma seleção rápida à informação necessária, porém como desafio, os sistemas de recuperação da informação devem coletar, representar, organizar e recuperar documentos considerando outras formas de representação da informação (MAIA, 2008).

Sintagmas são entendidos como grupos de palavras que fazem parte de sequências maiores adotados na estruturação de um texto, em que podem ou não ser facilmente identificados (SOUZA; ALVARENGA, 2004). Kuramoto (1996) em sua definição coloca os sintagmas nominais como "a menor parte do discurso portadora de informação".

> Parte-se da hipótese de que os sintagmas nominais, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos... (SOUZA; ALVARENGA, 2004)

Uma boa maneira de representar o conteúdo da internet é por meio do entendimento de formulários pelo computador e proceder com o uso de técnicas inteligentes para melhor aproveitar a representação (ANTONIOU, 2004). O homem, após descobrir uma nova tecnologia para transmissão e armazenamento de informação, demora um certo tempo para se adaptar e explorar os recursos disponíveis no novo meio (MAIA, 2008). Neste contexto os sintagmas nominais podem ser usados como meio de obter a informação necessária de forma eficaz.

3.2.4 Indexação de Textos

Os índices são utilizados para a recuperação rápida e precisa da informação. A forma como estes índices serão manipulados está ligado diretamente à tecnologia empregada (MAIA, 2008). Com o avanço tecnológico processadores mais ágeis podem ser usados para realizar as iterações necessárias para a extração dos índices de um documento eletrônico.

O processo de indexação visa a produção de uma lista de descritores por meio da extração das informações relevantes presentes em um documento (MAIA, 2008). A maioria dos métodos

ligados à recuperação da informação utilizam a palavra como unidade básica de acesso à informação (KURAMOTO, 2002). Logo em muitos métodos as palavras também são utilizadas como unidade lógica de indexação.

A utilização de palavras como descritores da informação passa pelos seguintes problemas destacados por Kuramoto (2002):

a) Polissemia: uma palavra com vários significados;
b) Sinonímia: duas palavras com o mesmo significado;
c) Combinação de duas ou mais palavras podem designar diferentes significados.

Para solucionar estes problemas é proposta a utilização de SN como descritores.

Alguns modelos já foram propostos para uma indexação automática dos descritores bem sucedida. O modelo boleano é um dos mais usados na recuperação da informação, se baseia na álgebra de Boole, em que os termos de indexação são combinados com operadores booleanos (and, or e/ou not). Porém este modelo apresenta pouca precisão nos resultados obtidos (KURAMOTO, 2002). Já o método Vetorial proposto por Baeza-Yates e Berthier (1999) se baseia na comparação parcial entre a representação dos documentos e a consulta do usuário. Pesos são atribuídos aos termos indexados, permitindo o cálculo do grau de similaridade entre um documento e os parâmetros de busca, o que evidencia a relevância de cada documento em relação à consulta.

Enquanto o modelo Booleano atribui pesos binários, ou 0 ou 1, o modelo Vetorial atribui um peso que pode variar entre 0 e 1, no universo dos números reais. Existem também modelos que utilizam a análise de referência, ou *link analysis*, em que os documentos passam a incorporar *hyperlinks* de outros documentos, gerando assim uma estrutura textual (KURAMOTO, 2002).

3.3 Similaridade de Documentos

A classificação presente nas atividades cotidianas já exemplifica a similaridade utilizada no processo de associação humano. Em seu trabalho, Maia (2008) verifica os aprimoramentos em medidas de similaridade entre documentos eletrônicos por meio de técnicas de processamento de linguagens naturais propostas por estudos que envolvem a ciência da informação, a computação e a linguística.

Quanto maior o número de características submetidas para análise de similaridade, mais confiável se torna o grau de similaridade entre os documentos (WIVES, 1999). O modelo proposto por Maia (2008) apresenta um enorme avanço no que diz respeito a extração automatizada e análise das características textuais dos documentos.

Por meio das propriedades de um objeto é possível a classificação automática das classes a qual o objeto está vinculado. Com esta classificação, dois ou mais documentos podemos possuir classes em comum, apresentando assim um nível de similaridade.

Em processos de classificação manual de documentos é são utilizadas espécies de etiquetas que podem apresentar por exemplo o assunto, ou o período que o documento foi escrito. Já em processos automatizados, algoritmos e técnicas computacionais buscam resultados semelhantes quanto à classificação e a etiquetação de documentos. O volume informacional neste contexto amplia a importância de pesquisas voltadas para a classificação de documentos eletrônicos de forma automatizada (MAIA, 2008). Já que consiste em processos rápidos de análise, mesmo que o tempo de processamento computacional ainda seja um limitador para análise de múltiplos documentos simultaneamente.

Estes algoritmos que resultam no valor do grau de similaridade entre documentos utilizam métricas que revelam o quanto os documentos se assemelham (MAIA, 2008). A medida básica utilizada no estudo de similaridade utiliza o cálculo do cosseno entre os ângulos formados por documentos em um espaço vetorial. O valor resultante é dado no universo dos números reais e representa o grau de correlação entre os documentos. A total semelhança entre documentos, ou seja, documentos idênticos, apresentam 1 como grau de similaridade. Já valores próximos de 0, ou 0 representam pouca semelhança entre, ou nenhuma relação entre os documentos.

A distância euclidiana entre os documentos O cálculo da distância entre estes documentos é dado por:

$$Distancia(p1,p2) = \sqrt{\sum_i (pesoTermo_{i1} - pesoTermo_{i2})^2}$$

É possível destacar algumas considerações acerca da distância, como a distância entre dois documentos deve ser maior ou igual a 0; a distância entre dois documentos idênticos é 0; a distância entre o documento 1 para o documento 2 é igual a distância entre o documento 2 para o documento 1 (MAIA, 2008).

3.4 Agrupamento de Documentos

O método utilizado neste projeto com a finalidade de agrupar os documentos é comumente usado em estudos de divergência genética no ramo da biologia agrícola, para auxiliar a classificação e na quantificação da distância genética entre fenótipos de plantas. Este trabalho fez uma adaptação do método de agrupamento por otimização, ou método de Tocher, apresentado por VASCONCELOS et al. (2007).

"Os métodos de agrupamento têm por finalidade separar um grupo original de observações em vários subgrupos, de forma a obter homogeneidade dentro e heterogeneidade entre os subgrupos." BERTAN, 2006.

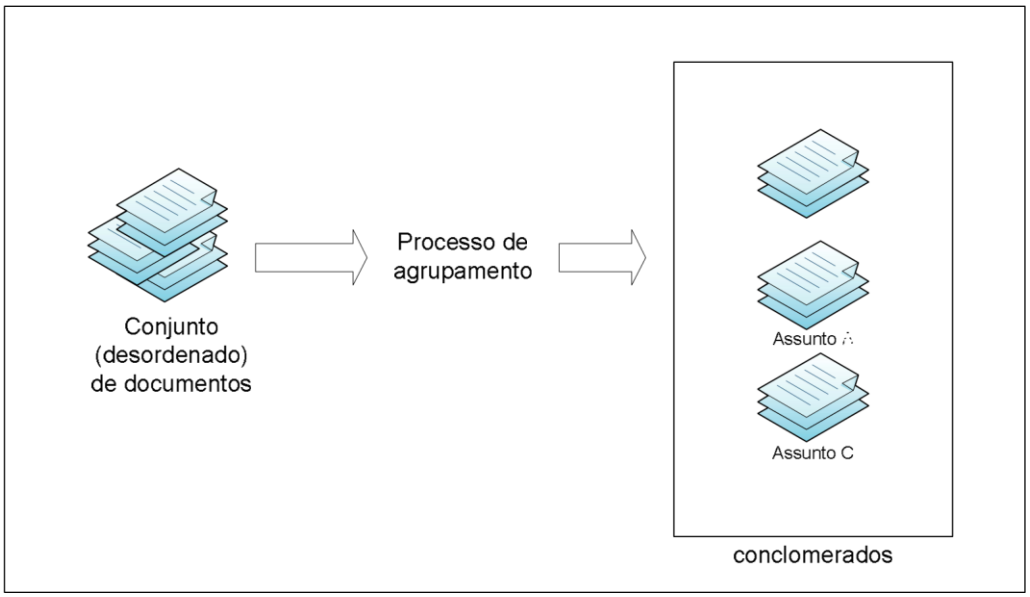

FIGURA 5 - Agrupamento de documentos por assunto
Fonte: Maia, 2008.

O tempo necessário para o agrupamento de qualidade está diretamente relacionado ao número de documentos e características submetidas à ferramenta de análise (WIVES, 1999). Neste contexto os termos e sintagmas nominais são as características que são passadas para o cálculo

de similaridade. Nesta etapa é levantada a distância euclidiana entre um documento perante todos os outros selecionados para serem agrupados.

Como se trata de distância, os valores calculados do cosseno de similaridade foram invertidos. De modo que se o cosseno de similaridade entre dois documentos é igual a 1 (textos iguais), a distância entre eles será 0. O mesmo acontece para os valores que estão nesta faixa, por exemplo documentos similares, podem apresentar cerca de 0,8 no cosseno de similaridade, a distância entre eles logo será 0,2.

O método de Tocher se baseia na identificação do texto mais similar a cada documento de um universo de textos. Formando assim pares, com a distância entre eles menor que a distância entre outros elementos. Após a identificação, é então formado o primeiro grupo, com o par de documentos mais próximo entre si. Também é considerada a maior distância entre os pares, como α. A distância entre um elemento e outro do mesmo grupo não pode ser maior que α, de forma a manter a distância média intragrupo sempre inferior a qualquer distância intergrupo (BERTAN, 2006).

Como critério para o agrupamento, é levantado a distância média entre três pontos. No caso do primeiro grupo, dois pontos já foram levantados, em seguida é calculado a distância média entre o par e os demais documentos, por meio da expressão:

$$D3pts = \quad (Distancia(p1,p3) + Distancia(p2,p3)) - Distancia(p1,p2)$$

Se o resultado obtido por D3pts for menor que α a inclusão do elemento 3 no grupo é permitida, caso contrário o elemento é considerado de outro grupo e será agrupado posteriormente.

Após a validação dos elementos no primeiro um grupo, novamente, são formados os pares com os documentos mais próximos de cada grupo, desconsiderando os elementos já inseridos em algum grupo. Um novo grupo é formado com o novo par encontrado, e outros elementos se unem a este grupo por meio da mesma validação realizada na primeira execução. O processo acontece até que todos elementos sejam agrupados.

4. METODOLOGIA

O modo como os usuários buscam a informação na web possui características peculiares, uma vez que a busca normalmente acontece por meio dos mecanismos de buscas e os usuários procuram por uma ou duas palavras chaves, em um domínio de bilhões de páginas web, perante este cenário uma busca bem sucedida está relacionada ao resultado apresentado pela consulta e se este resultado é considerado como o resultado "esperado" (NIELSEN; LORANGER, 2007). A presente pesquisa está focada na abordagem dos sintagmas nominais como forma de classificação de documentos eletrônicos visando encontrar similaridades entre eles.

> "Os métodos de descoberta de conglomerados ou classificação se mostraram extremamente dependentes de técnicas de pré-processamento dos textos que visassem a padronizá-los, minimizando os problemas do vocabulário e representando seu conteúdo de forma mais correta e fácil de ser trabalhada pela máquina." (MAIA, 2008).

A utilização dos sintagmas nominais como descritores de informação demonstram imensos ganhos para o entendimento sistemático da língua portuguesa. Ainda assim, não é possível encontrar muitas aplicações com estas características disponível na internet. A interface sugerida pela aplicação implementada por MAIA (2008) e adaptada neste trabalho representa um avanço, mas não possui pode ser entendida como uma solução definitiva para recuperação de informações contidas em documentos. O agrupamento aqui sugerido também não simboliza uma solução definitiva para o processo de catalogação de documentos. A pesquisa busca demonstrar a eficiência do processo de classificação da informação por meio da análise automática de documentos textuais com base nos SN.

4.1 Etapas da Pesquisa

A pesquisa se desenvolve nas seguintes etapas:

- Adaptação da ferramenta OGMA para plataforma web.
- Aplicação da metodologia de agrupamento de elementos com base na distância euclidiana.
- Importação do corpus
- Análise Textual, extração dos SN
- Agrupamento Automatizado do Corpus

4.1.1 OGMA Web

Para realização da pesquisa foi necessária a adequação da aplicação OGMA para uma ferramenta voltada à plataforma web, devido à facilidade de acesso que a plataforma

disponibiliza. O aplicativo foi desenvolvido na linguagem PHP (Hypertext Processor), devido à grande utilização da linguagem para ferramentas web e pode ser acessado pelo endereço web: www.ogmaweb.com.br.

O primeiro desafio para a adaptação da ferramenta foi a compreensão das funcionalidades do OGMA em se tratando de extração dos SN, análises de texto. A conversão do código da linguagem C# para a linguagem web responsável pela automatização do processo seguiu de forma conjunta com a incorporação do banco de dados convertido de Acess para MySQL.

Por se tratar de uma aplicação web, foi incorporada a funcionalidade de identificação dos usuários, com a finalidade de atribuir um acervo personalizado para cada usuário cadastrado na ferramenta, de maneira a possibilitar a livre construção de bases de documentos eletrônicos para análises automatizadas. A Figura 9 apresenta os recursos existentes na ferramenta OGMA que foram mantidos na plataforma web.

FIGURA 6 – Funcionamento da ferramenta

As funcionalidades existentes no OGMA como a extração de termos e a atribuição de pesos vinculados à frequência de aparição no texto; a consideração à lista de stopwords; a extração dos sintagmas nominais, sintagmas nominais únicos e pontuados, o cálculo de similaridade e o método de etiquetação foram mantidas na ferramenta web.

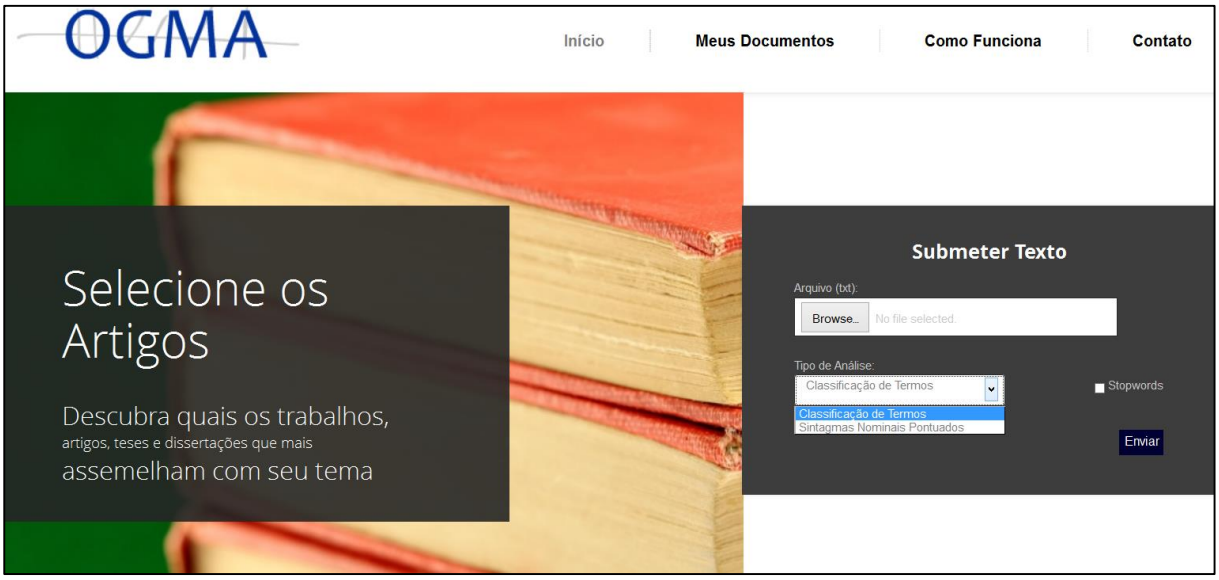

FIGURA 7 – Tela de Inserção de Documentos do OGMA Web

Após a submissão de um documento a ferramenta disponibiliza três colunas de resultados, conforme os parâmetros que foram passados para o processamento.

Na primeira coluna são exibidos três ícones que nos levam à três documentos resultantes da análise do texto submetido. O texto etiquetado, os sintagmas nominais, ou termos de acordo com a seleção do tipo de análise realizada na tela de inserção e a classificação destes sintagmas ou termos levando em consideração a ponderação adotada e o método de extração selecionado pelo usuário. Também são exibido os documentos já inseridos pelo usuário, para a realização de cálculos de similaridade entre os documentos.

Na segunda coluna são exibidos os 30 sintagmas nominais ou termos que apresentaram maior relevância no texto, já em ordem classificatória, com o número de ocorrências ao lado.

A terceira coluna relaciona os resultados encontrados com uma consulta no Google Scholar (http://scholar.google.com) por meio de uma busca utilizando os três primeiros sintagmas ou termos encontrados. O link Mais Resultados na parte inferior da coluna nos leva à página do Google que disponibilizou os dados.

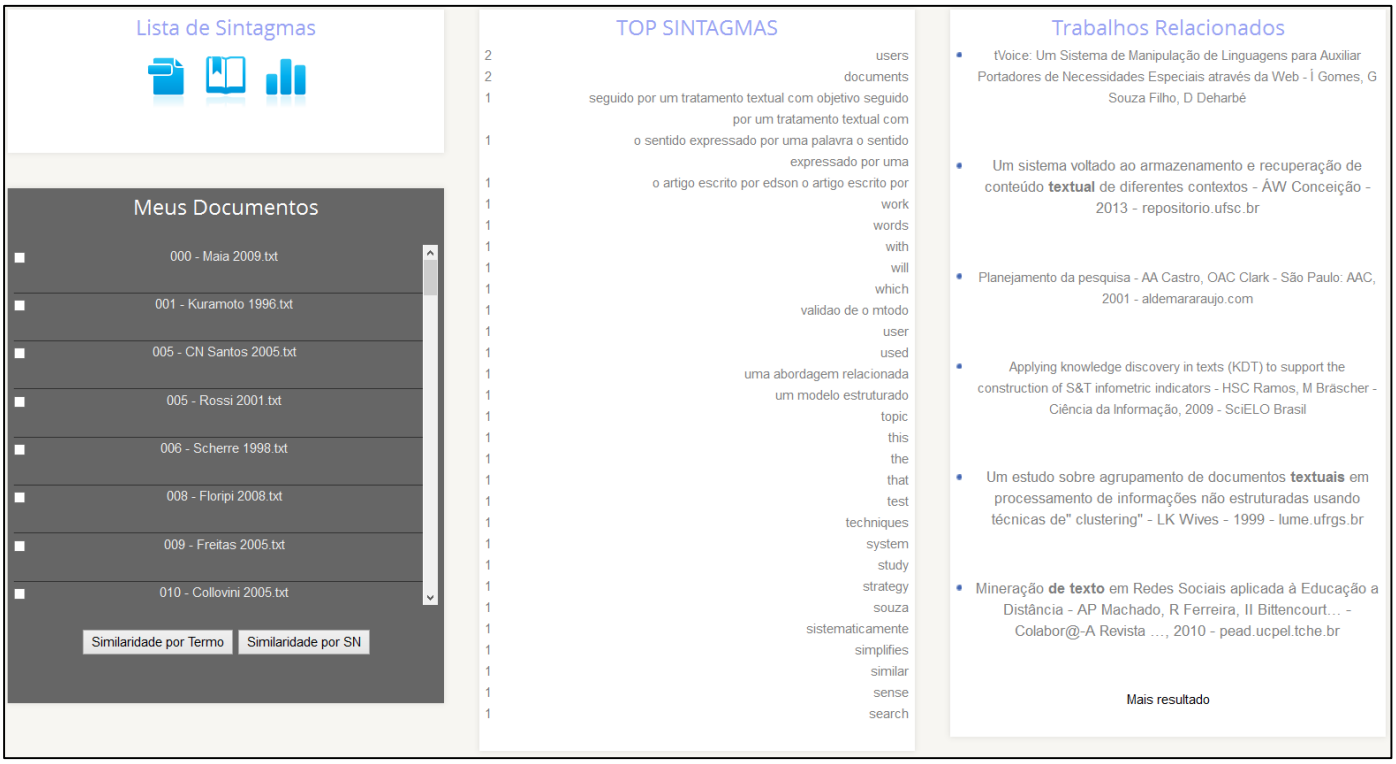

FIGURA 8 - Tela de Resultados do OGMA Web

A página Meus Documentos nos leva a listagem dos documentos já enviados pelo usuário e nos permite a gestão destes documentos, bem como o cálculo de similaridade. Ao clicarmos no nome do documento é exibido as primeiras informações contidas nele. Com a seleção de dois documentos o OGMA Web pode compara o cosseno de similaridade, que representa o ângulo entre dois vetores em um espaço vetorial (MAIA, 2008) e após submissão para o cálculo é retornado um valor entre 0 e 1. Em que 0 significa que nenhuma relação foi encontrada entre estes documentos, e 1 para documentos com total similaridade.

FIGURA 9 - Tela de Documentos do OGMA Web
Fonte: Elaborada pelo autor da dissertação.

O resultado do cálculo do consenso, pode ser visto na FIGURA 4, em que os trabalhos de Hélio Kuramoto (1996) e Luiz Maia (2008) foram comparados.

FIGURA 10 - Resultado do cálculo de similaridade
Fonte: Elaborada pelo autor da dissertação.

Caso mais de dois documentos sejam selecionados para o cálculo de similaridade a funcionalidade responsável pelo agrupamento é acionada. A primeira etapa consiste no levantamento da distância entre os documentos selecionados entre si. Com base na matriz na matriz de distância os elementos são então agrupados seguindo o método de Tocher e em seguida, com o intuito de identificar um descritor para o grupo é pesquisado o termo ou o sintagma nominal mais relevante.

FIGURA 11 – Resultado do agrupamento
Fonte: Elaborada pelo autor da dissertação.

No exemplo foram coletados artigos de um site de notícias disponível na internet de 2 assuntos diferentes, futebol e política. Também foram submetidos 2 resenhas que abordam sintagmas nominais. Como resultado, a ferramenta apresentou os grupos já identificados e os textos pertencentes a cada grupo. Também é apresentado um link que nos leva a matriz de distância utilizada no cálculo.

O banco de dados foi convertido para MySQL e novas mudanças ocorreram na estruturação dos dados. Foram criadas associações entre os usuários e seus documentos para que os acervos sejam personalizados. Também foram armazenados os termos do documento processado para uma futura análise de similaridade, o que reduz o tempo de resposta da ferramenta.

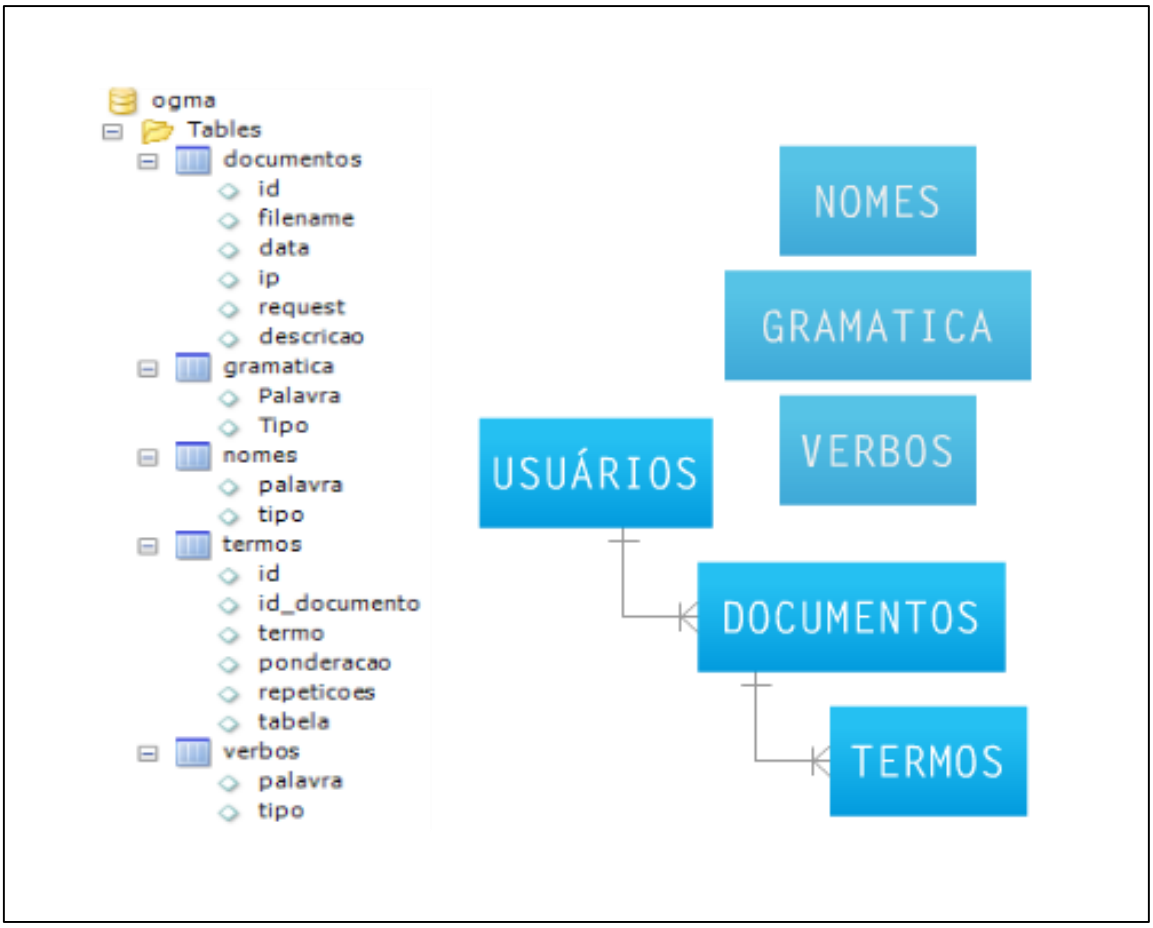

FIGURA 12 – Estruturação dos dados
Fonte: Elaborada pelo autor da dissertação.

4.1.2 Validação da Ferramenta

A validação do OGMA Web se dividiu em etapas. Uma vez que a ferramenta foi adaptada da plataforma Microsoft Windows para um ambiente distribuído, a primeira etapa consistiu na comparação dos dados retornados pelas duas ferramentas, OGMA e OGMA Web.

Foram utilizados os experimentos de Luiz Maia (2008), em que foi verificado os resultados alcançados com a extração automática de SN, realizada pelo OGMA, em comparação com os SN extraídos por outra ferramenta textual, o VISL (*visual interactive syntax learning*). No experimento foram selecionados alguns textos para aprimoramento e ajustes das regras feitas automaticamente pelo OGMA.

Os 6 textos encontrados no ANEXO I, do estudo de Maia (2008), foram submetidos ao OGMA Web para assegurar a conformidade entre os resultados encontrados pelas ferramentas. Quatro

destes textos obtiveram a mesma listagem de SN nas duas plataformas, OGMA e OGMA Web. Em alguns casos a ordem dos sintagmas encontrados não foi igual, o que não representa risco. Em um dos arquivos comparados houve uma inconsistência entre os resultados. Os resultados da comparação estão no ANEXO II deste estudo.

Em outro caso, a ferramenta OGMA identificou o SN: “suécia e vaticano”, e a ferramenta web identificou apenas: “suécia”. Com a análise realizada no texto etiquetado, foi constatado que havia diferenças no vocabulário utilizados pelas ferramentas. A palavra “vaticano” no OGMA Web não estava definido como substantivo e no OGMA havia esta caracterização. Após a correção da classificação da palavra no vocabulário os resultados foram o mesmo.

Na última inconsistência encontrada, ambas ferramentas identificaram o sintagma: “artigos aindanão publicados”. As palavras “ainda” e “não” se juntaram nos SN encontrados nas duas ferramentas. O que pode representar um problema. Em seguida, apenas o OGMA Web identificou os SN: “ainda não publicados”, “projeto internacional”. Como o OGMA Web apresentou resultados mais convincentes, a incoerência entre os resultados aqui descritas não serão levadas a diante.

Após a validação da extração dos SN, o cálculo de similaridade também foi testado para assegurar o correto funcionamento da ferramenta web. Considerando as regras de correlação de Pearson, os valores obtidos na similaridade entre dois documentos não podem ser maior que 1, nem menor que 0. O cálculo do cosseno de similaridade acontece para medir o ângulo formado entre dois elementos em um espaço vetorial, com isso documentos semelhantes tendem possuir um fator de similaridade próximo de 1. Para a validação foram submetidos 4 documentos, sendo 2 com o mesmo texto, o que implica no cosseno de similaridade 1, se comparados. Um dos documentos continha apenas palavras do idioma inglês. O último documento contém um conteúdo similar ao primeiro (e ao segundo).

A similaridade entre os dois documentos por meio dos sintagmas ocorreu como esperado, os documentos iguais apresentaram 0 como cosseno de similaridade e quando comparados ao documento 3 (texto em inglês) retornaram 0. A comparação entre eles e o documento 4 ocorreu de duas formas, entre sintagmas nominais pontuados e termos com stopwords. Os resultados obtidos foram respectivamente 0.003403536203889 e 0.50467534129741. Como se trata de

documentos semelhantes acerca de um determinado tema, sintagmas nominais pontuados podem não ser a melhor opção para medir a similaridade entre documentos, apesar de apresentarem enorme eficiência na coleta de descritores portadores de informações relevantes.

4.1.3 Aplicação da Ferramenta no Corpus

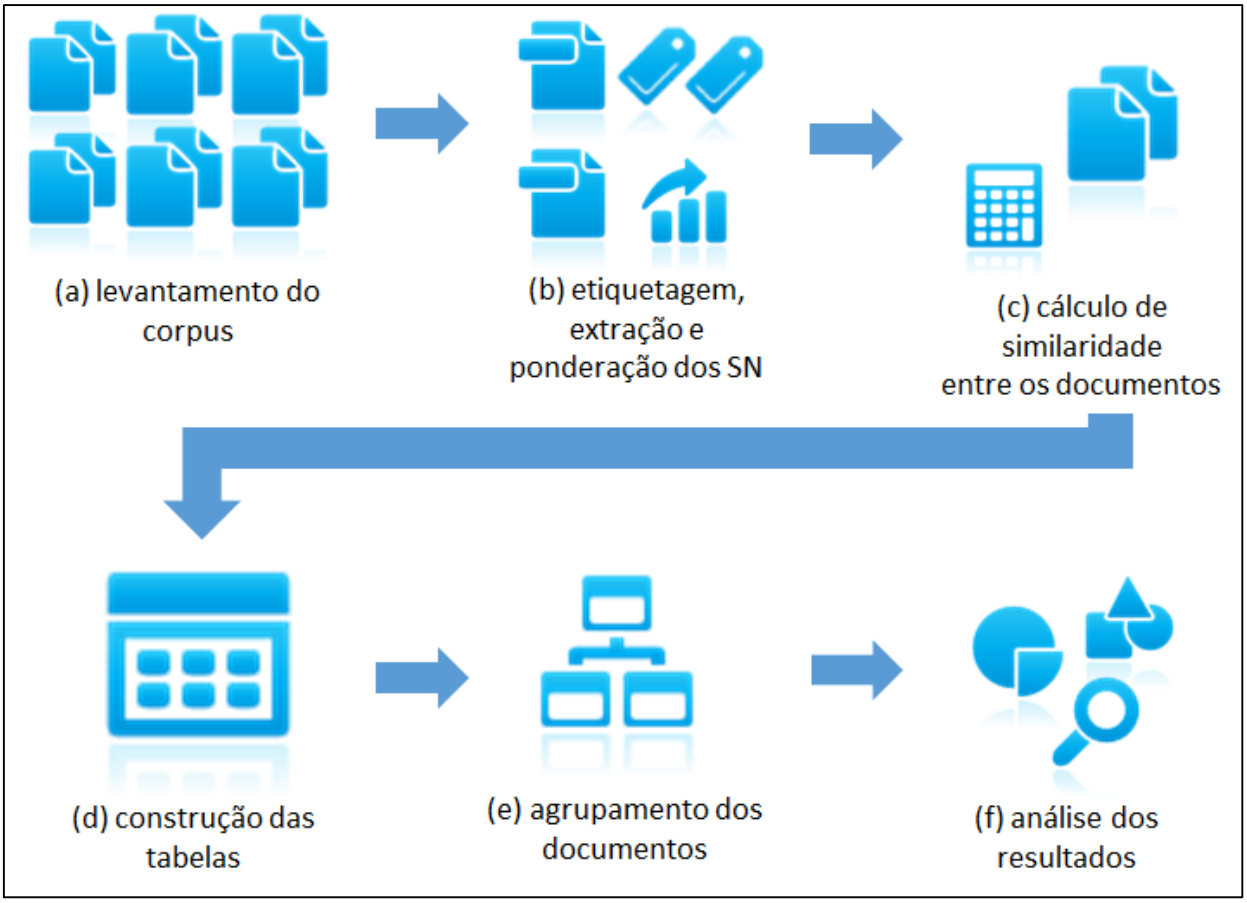

FIGURA 13 – Etapas do desenvolvimento
Fonte: Elaborada pelo autor da dissertação.

a) Levantamento do corpus.
b) Etiquetagem, extração e ponderação dos termos.
c) Cálculo de similaridade entre os documentos.
d) Construção da tabela de distância entre os documentos.
e) Agrupamento dos documentos.
f) Análise dos resultados.

O experimento será aplicado sobre um corpus formado por 12.011 resumos de artigos, extraídos de forma automática do site da Associação Nacional de Pós-Graduação e Pesquisas em Administração (ANPAD), por meio de um extrator de dados dinâmico. Muitos destes trabalhos foram publicados em edições do encontro da associação, o EnANPAD, durante os anos 1997 e 2012.

A estrutura dos resumos extraídos é formado pelo título, o evento ao qual o trabalho foi publicado, o texto do resumo e também a área de classificação do estudo, que representa uma categorização, consequentemente um agrupamento prévio dos trabalhos. Foram constatadas 272 áreas diferentes no corpus.

Após a extração dos resumos, de forma automática, os dados foram lapidados para serem analisados pelo OGMA Web. Como se trata de um grande acervo, os processos de análise textual, cálculos de similaridade e agrupamento podem apresentar bastante impacto no tempo de processamento dos dados, o que impede a utilização da interface básica de trabalho disponível pelo OGMA Web.

Todos os textos dos resumos passaram pelo processo de classificação de SN por meio de uma rotina recursiva responsável pela solicitação. Em seguida serão elaborados outros procedimentos recursivos com a finalidade de calcular as distâncias entre um documento perante todos os outros pertencentes ao corpus com base no cálculo de similaridade proposto por Maia (2008). Formando assim uma tabela de distância, utilizada como insumo para o agrupamento automático dos documentos.

Com o agrupamento será possível novas análises acerca do tema central de cada grupo, afim de identificar um possível descritor, ou título para o grupo em questão.

A divisão realizada automática será comparada com a classificação prévia do corpus com o intuito de determinar a acurácia do método. Como a quantidade de grupos é definida dinamicamente conforme as regras do método de Tocher, é possível que sejam encontrados um número distinto de grupos entre a classificação automática e a pré-classificação do corpus, o que pode demandar uma nova lapidação da classificação já existente.

O resultado esperado pretende assegurar a eficiência do método de agrupamento automático de documentos eletrônicos, por meio da análise dos grupos apresentados pela ferramenta. Os grupos ao serem formados podem simbolizar um avanço no tratamento de informação, principalmente no que diz respeito à catalogação da informação.

5. CRONOGRAMA

Para que os objetivos da pesquisa sejam alcançados foram separadas as atividades necessárias para a elaboração do projeto e definido o intervalo de tempo e o ponto de início de cada tarefa.

REFERÊNCIAS

ALVARENGA, L. A teoria do conceito revisitada em conexão com ontologias e metadados no contexto das bibliotecas tradicionais e digitais. **DataGramaZero–Revista de Ciência da Informação,** v. 2, n. 6, 2001.

_____. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli**, n. 15, 2003. ISSN 1518-2924.

ANTONIOU, G. **A semantic web primer**. the MIT Press, 2004. ISBN 0262012103.

BASKERVILLE, R.; DULIPOVICI, A. The theoretical foundations of knowledge management. **Knowledge Management Research & Practice,** v. 4, n. 2, p. 83-105, 2006. ISSN 1477-8238.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american,** v. 284, n. 5, p. 28-37, 2001. ISSN 0036-8733.

CASTELLS, M. A galáxia da internet: reflexões sobre a internet, os negócios ea sociedade; tradução Maria Luiza X. de A. **Borges. Rio de Janeiro: Jorge Zahar**, 2003.

CHOO, C. W. **A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões**. Senac, 2006. ISBN 8573593415.

DANTAS, M. Capitalismo na era das redes: trabalho, informação e valor no ciclo da comunicação produtiva. **Informação e globalização na era do conhecimento. Rio de Janeiro: Campus**, p. 216-261, 1999.

DE VASCONCELOS, E. S. et al. Método alternativo para análise de agrupamento. **Pesquisa Agropecuária Brasileira,** v. 42, n. 10, p. 1421-1428, 2007. ISSN 1678-3921.

DIAS, E. J. W. Biblioteconomia e ciência da informação: natureza e relações. **Perspectivas em Ciência da Informação,** v. 5, 2000. ISSN 1981-5344.

DIAS, E. W. Contexto digital e tratamento da informação. **DataGramaZero-Revista de Ciência da Informação,** v. 2, n. 5, 2001.

GARCÍA JIMÉNEZ, A. Instrumentos de representación del conocimiento: tesauros versus ontologías. Anales de documentación, 2004. p.79-95.

KURAMOTO, H. Uma abordagem alternativa para o tratamento ea recuperação de informação textual: os sintagmas nominais. **Ciência da informação,** v. 25, n. 2, 1996. ISSN 1518-8353.

_____. Sintagmas nominais: uma nova proposta para a recuperação de informação. 2002. ISSN 1517-3801.

LANCASTER, F. W. **Toward paperless information systems**. Academic Press, Inc., 1978. ISBN 0124360505.

LANCASTER, F. W.; WARNER, A. J. **Information Retrieval Today. Revised, Retitled**. ERIC, 1993. ISBN 0878150641.

MAIA, L. C.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação,** v. 15, p. 154-172, 2010. ISSN 1413-9936. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362010000100009&nrm=iso99362010000100009&nrm=iso >.

MAIA, L. C. G.; SOUZA, R. R. MEDIDAS DE SIMILARIDADE EM DOCUMENTOSELETRÔNICOS. **Escola de Ciência da Informação, UFMG. Artigo apresentado no IX ENANCIB–Encontro Nacional de Pesquisa em Ciência da Informação, USP**, 2008.

MARCO, G.; ESTEBAN NAVARRO, F. On some contributions of the cognitive sciences and epistemology to a theory of classification. **Knowledge organization,** v. 20, n. 3, p. 126132, 1993. ISSN 0943-7444.

MOREIRA, A.; ALVARENGA, L.; OLIVEIRA, A. D. P. O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. **DataGramaZero-Revista de Ciência da Informação,** v. 5, n. 6, 2004.

PERINI, M. A. **A gramática gerativa: introdução ao estudo da sintaxe portuguesa**. Editora Vigília, 1976.

_____. **Sofrendo a gramática: ensaios sobre a linguagem**. Editora Atica, 2002. ISBN 8508067291.

PETERSEN, K. et al. **Systematic mapping studies in software engineering**. 12th International Conference on Evaluation and Assessment in Software Engineering, 2008.

SETZER, V. W. Dado, informação, conhecimento e competência. **DataGramaZero Revista de Ciência da Informação, n. 0**, 1999.

SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. **Perspectivas em Ciência da Informação,** v. 10, n. 2, 2008. ISSN 1981-5344.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação, Brasília,** v. 33, n. 1, p. 132-141, 2004.

STAAB, S.; STUDER, R. **Handbook on ontologies**. Springer, 2009. ISBN 3540926739.

STOJANOVIC, L.; STAAB, S.; STUDER, R. eLearning based on the Semantic Web. WebNet2001-World Conference on the WWW and Internet, 2001. p.23-27.

VICKERY, B. C. Knowledge representation: a brief review. **Journal of documentation,** v. 42, n. 3, p. 145-159, 1986. ISSN 0022-0418.

ANEXOS

I. Revisão Sistemática da Literatura: Trabalhos encontrados

Nome
Uso de sintagmas nominais na classificação automática de documentos eletrônicos - LC Maia, RR Souza 2010
Uma abordagem alternativa para o tratamento ea recuperação de informação textual: os sintagmas nominais - H Kuramoto - 2002
Extração de sintagmas nominais para o processamento de co-referência - R Vieira, R CHISHMAN, F GORZIZA, R ROSSONI 2000
Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais - RR Souza – 2005
Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro - CN dos Santos 2006
Resolução de correferência em textos da língua portuguesa - D Rossi, C Pinheiro, N Feier, R Vieira 2001
Sobre a concordância de número no português falado do Brasil - MMP Scherre, AJ Naro - Dialettologia, geolinguistica, sociolinguistica, 1998 - ai.mit.edu
Estudo da variação do determinante em sintagmas nominais possessivos na história do português - SA Floripi - 2008 - bibliotecadigital.unicamp.br
A anotaçao de um corpus para o aprendizado supervisionado de um modelo de sn - MC de Freitas, M Uzeda-Garrão, C Oliveira… - Proceedings of the III …, 2005 - lbd.dcc.ufmg.br
Classificação automática de expressões anafóricas em textos da língua portuguesa - S Collovini, JCB Coelho, R Vieira - Proceedings of ENIA, 2005 - lbd.dcc.ufmg.br
Língua de sinais na educação dos surdos - LB Karnopp - A invenção da surdez: cultura, alteridade, identidade e …, 2004
Uma ferramenta para resolução automática de correferência - C Gasperin, R Goulart, R Vieira - IV Encontro Nacional de …, 2003 - rodrigo.goulart.nom.br
Um fragmento da constituição sócio-histórica do Português do Brasil: Variação na concordância nominal de número em um dialeto Afro-brasileiro - PR Andrade - 2003 - Dissertação de mestrado
DEBATE/DEBATE AS DUAS GRANDES VERTENTES DA HISTÓRIA SOCIOLINGÜÍSTICA DO BRASIL (1500-2000) - D Lucchesi - Delta, 2001 - SciELO Brasil
Origens do português brasileiro - AJ Naro, MMP Scherre - 2007 - Parabola Books
Restrições sintáticas e semânticas no controle da concordância verbal em português - MMP Scherre, AJ Naro - Fórum lingüístico, 2008 - 150.162.1.115
Análise de textos orais - D Preti, ACS Rodrigues - 1993 - FFLCH/USP, Projeto de Estudo da …
Um corpus de textos científicos em Português para a análise da Estrutura Esquemática - VD Feltrim, MGV Nunes, SM Aluísio - Série de Relatórios do Núcleo Interinstitucional …, 2001
A temática da cidadania na imprensa escrita de São Paulo: análise lexical e sociossemiótica - RMN Prados - 2000
TeMário: Um Corpus para Sumarização Automática de Textos - TAS Pardo, LHM Rino - São Carlos: Universidade de São Carlos, …, 2003 - icmc.usp.br
A noção de formação discursiva: uma relação estreita com o corpus na análise do discurso - VMO SARGENTINI - … DE ESTUDOS EM ANÁLISE DO …, 2005 - analisedodiscurso.ufrgs.br
Anáforas associativas actanciais e nominalizações: delimitação do ponto de vista da semântica de eventos - G Zamponi - Cadernos de estudos lingüísticos, 2011 - espea.iel.unicamp.br
Terminologia revisitada - M da Graça KRIEGER - Delta, 2000 - SciELO Brasil
A lingüística ea ciência da informação: estudos de uma interseção - ES Mendonça - Ciência da Informação, Brasília, 2000 - SciELO Brasil
Manuais acadêmicos de Química Geral em língua portuguesa: aspectos lingüístico-terminológicos e aspectos conceituais - MJB Finatto, N Enzweiler, C Huang, M Eichler… - TradTerm, 2002 - iq.ufrgs.br
Curso Básico de Terminologia Vol. 54 - LA Barros - 2004 - books.google.com
As ciências do léxico - AN Isquerdo, MG KRIEGER - … , Lexicografia, Terminologia. …, 2010 - geltra.ibilce.unesp.br

Escolhas estratégicas e estilo cognitivo: um estudo com pequenas empresas - FAP Gimenez - Revista de Administração Contemporânea, 1998 - SciELO Brasil
Identificação dos critérios de avaliação de resultados do serviço de enfermagem nos programas de acreditação hospitalar - SDEEENP DE - Rev Latino-am Enfermagem, 2006 - sjc.unifesp.br
Extração de conhecimento e análise visual de redes sociais - …, R Galante, LC Lamb, AS Spritzer… - … Integrado de Software …, 2008 - sitedaescola.com
Desigualdade salarial: resultados de pesquisas recentes - JGA Reis, RP BARROS - Distribuição de renda no Brasil. São Paulo: Paz e Terra, 1991
Psicologia das massas ea análise do eu - S Freud - Edição standard brasileira das obras …, 1921 - e.livros.clube-de-leituras.pt
Método para la resolución de correferencias de sintagmas nominales dfinidos incluyendo alias y acrónimos en el sistema de información EXIT - A Ferrández, R Muñoz, PM Barco - Procesamiento del lenguaje …, 1999 - dialnet.unirioja.es
Coreference resolution between sources of opinions in Spanish texts - F Acerenza, M Rabosto, M Zubizarreta… - … (CLEI), 2012 XXXVIII …, 2012 - ieeexplore.ieee.org
Detecting anaphoricity and antecedenthood for coreference resolution - O Uryupina - María Teresa Vicente-Díez, Paloma Martínez, Ángel …, 2009 - sepln.org
Text as scene: Discourse deixis and bridging relations - M Recasens, MA Marti, M Taulé - Procesamiento del lenguaje natural, 2007 - sepln.org
Definite Descriptions Resolution in Spanish - R Munoz, A Ferrández, CSV del Raspeig - Proceeding of ACIDCA'2000: …, 2000 - Citesee
Variedades y serotipos de Cryptococcus neoformans en pacientes con SIDA y neurocriptococosis en São Paulo, Brasil - B Calvo, O Fischman, A Pignatari… - Revista do Instituto de …, 1990 - SciELO Brasil
Uso de los servicios de salud y progresión al Sida entre personas con infeccíon por VIH en Belo Horizonte (Minas Gerais), Brasil - FA Acurcio, MDC Guimarães - Revista Panamericana de …, 1998 - SciELO Public Health
Discursos a la nación mexicana: por Antonio Caso - A Caso - 1922 - Porrua hnos.
Clinical nurses attitude towards alcoholic patients - P Reyes Navarrete, MAV Luis - Revista latino-americana de …, 2004 - SciELO Brasil
Sistema experto de diagnóstico clínico para el apoyo de la primera consulta - J Rozenbojm, E Palladino, AC Azevedo - Salud Pública Mex, 1993
Los instrumentos normativos en ética de la investigación en seres humanos en América Latina: análisis de su potencial eficácia - C Lorenzo, G Keyeux, V Penchaszadeh, A Saada - … investigacíon en seres humanos …, 2006
Research on the web and text production: reflexions on the teaching of argumentative types of text at school - DB Braga, MA Moraes - Linguagem em (Dis) curso, 2009 - SciELO Brasil
RHeSumaRST: Um sumarizador automático de estruturas RST - ERM Seno, LHM Rino - 2005 - bdtd.ufscar.br
The scholar critique and the lecturer reflection - MG Sá, GL Moura - Cadernos EBAPE. BR, 2008 - SciELO Brasil
ExtraWeb: um sumarizador de documentos Web baseado em etiquetas HTML e ontologia - PP Silva - 2006 - bdtd.ufscar.br
DANTAS, MARCO AURÉLIO RIBEIRO - I COPPE, UII Título - 2007 - wwwp.coc.ufrj.br
UNIVERSIDADE FEDERAL DE SÃO CARLOS DEPARTAMENTO DE LETRAS PROGRAMA DE PÓS-GRADUAÇÃO EM LÍNGÜÍSTICA - TI Carbonel - 2007 - bdtd.ufscar.br
Subjetividade, argumentação, polifonia: a propaganda da Petrobrás - HHN Brandão - 1998 - Editora Unesp Fundac~ao
Terra à vista: discurso do confronto: velho e novo mundo - EP Orlandi - 1990 - Cortez Editora
UMA TEMPESTADE DE LUZ: A COMPREENSÃO POSSIBILITADA PELA ANÁLISE TEXTUAL DISCURSIVA A storm of light: comprehension made possible by … - R Moraes - Ciência & Educação, 2003 - SciELO Brasil
Aspectos do processamento do fluxo de informação no discurso oral dialogado - IGV Koch, C JUBRAN, H Urbano, LL Fávero… - Gramática do português …, 1990
Vozes e contrastes: discurso na cidade e no campo - EP Orlandi, E Guimarães, FL Tarallo - 1989 - Cortez Editora
O desafio de dizer não - S Lagazzy - 1988 - Pontes Editores

O objeto nulo no português brasileiro: percurso de uma pesquisa - CMC Galves - Cadernos de Estudos Lingüísticos, 2012 - iel.unicamp.br
As construções causativas em tapirapé - Y Leite - Revista Interamericana de Estudos Etnolingüísticos, 1994
Anáfora e correferência: por que as duas noções não se identificam? - R Ilari - Cadernos de Estudos Lingüísticos, 2011 - info03.iel.unicamp.br
Strutture testuali e retoriche - B Mortara Garavelli - Introduzione all'italiano contemporaneo. Le strutture, …, 1993
Le dislocazioni a destra in italiano - G Berruto - Tema-Rema in italiano, 1986 - books.google.com
Anáforas associativas actanciais e nominalizações: delimitação do ponto de vista da semântica de eventos - G Zamponi - Cadernos de estudos lingüísticos, 2011 - espea.iel.unicamp.br
Dicionário de teoria da narrativa - C Reis, ACM Lopes - 1988 - Editora Atica
Responsabilidade social empresarial e empresa sustentável: da teoria à prática - JC Barbieri, JER Cajazeira - 2009 - Ed. Saraiva
Literatura infantil: teoria, análise, didática - NN Coelho - 2000 - Moderna
Curso Básico de Terminologia Vol. 54 - LA Barros - 2004 - books.google.com
Ato/atividade e evento - A Sobral, B Brait - Bakhtin: conceitos-chave. São Paulo: Contexto, 2005
Teoria geral do federalismo - JA de Oliveira Baracho - 1982 - FUMARC/UCMG
Inteligência organizacional e competitiva - K Tarapanoff, JAR Gregolin - Ciência da Informação, 2002 - SciELO Brasil
A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões - CW Choo - 2006 - Senac
A ciência da informação como ciência social - CAÁ Araújo - Ciencia da informação, 2003 - SciELO Brasil
Introdução geral às ciências e técnicas da informação e documentação - C Guinchat, M Menou - 1994 - livroaberto.ibict.br
Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação ea difusão do conhecimento - NAP Vanti - Ciência da Informação, 2002 - SciELO Brasil
Ciência da informação: origem, evolução e relações - T Saracevic - Perspectivas em ciência da informação, 2008 - portaldeperiodicos.eci.ufmg.br
O método de pesquisa survey - H Freitas, M Oliveira, AZ Saccol, J Moscarola - Revista de Administração, 2000 - unisc.br
Métodos de pesquisa em administração - SC Vergara - 2005 - Atlas São Paulo
Avaliação da satisfação do consumidor utilizando o método de equações estruturais: um modelo aplicado ao setor elétrico brasileiro - R Marchetti, PHM Prado - Revista de Administração Contemporânea, 2004 - SciELO Brasil
Consumidores satisfeitos, e então? Analisando a satisfação como antecedente da lealdade - JA Larán, FS Espinoza - Revista de Administração Contemporânea, 2004 - SciELO Brasil
Construção de indicadores para avaliação de conceitos intangíveis em sistemas produtivos - M Sellitto, J Ribeiro - Gestão & Produção, 2004 - SciELO Brasil
Um guia para avaliação de artigos de pesquisas em sistemas de informação - N Hoppen, L Lapointe, E Moreau - … . Porto Alegre. Edição 3, vol. 2, n. 2 …, 1996 - lume.ufrgs.br

II. Validação: Resultados OGMA x OGMA Web

RESUMO 1: *SOUZA, Terezinha Batista, CATARINO, Maria Elisabete, SANTOS, Paulo Cesar dos. Metadados: catalogando dados na Internet. Transinformacao, Campinas, v. 9, n. 2, p. 93-105, maio/ago. 1997*

TEXTO: Apresenta de forma introdutória questões e conceitos fundamentais sobre metadados e a estruturação da descrição padronizada de documentos eletrônicos. Discorre sobre os elementos propostos no Dublin Core e comenta os projetos de catalogação dos recursos da Internet, CATRIONA, InterCat e CALCO.

OGMA	OGMA WEB
1. forma introdutória 2. introdutória questões e conceitos fundamentais sobre metadados e a estruturação da descrição padronizada de documentos eletrônicos 3. os elementos propostos no dublin 4. os projetos de catalogação dos recursos da internet 5. catriona 6. intercat	1. forma introdutória 2. introdutória questões e conceitos fundamentais sobre metadados e a estruturação da descrição padronizada de documentos eletrônicos 3. os elementos propostos no Dublin 4. os projetos de catalogação dos recursos da internet 5. catriona 6. intercat

RESUMO 2: *CUNHA, Murilo Bastos da. Biblioteca digital: bibliografia internacional anotada. Ciência da Informação, Brasília, v.26, n.2, p.195-213, maio/ago. 1997* .

TEXTO: Bibliografia internacional seletiva e anotada sobre bibliotecas digitais. Aborda os seguintes aspectos: a) Visionários, principais autores que escreveram sobre a biblioteca do futuro, no período de 1945-1985; b) conceituação de biblioteca digital; c) projetos em andamento na Alemanha, Austrália, Brasil, Canadá, Dinamarca, Espanha, Estados Unidos, Franca, Holanda, Japão, Nova Zelândia, Reino Unido, Suécia e Vaticano ; d) aspectos técnicos relativos a construção de uma biblioteca digital: arquitetura do sistema, conversão de dados e escaneamento, marcação de textos, desenvolvimento de coleções, catalogação, classificação/indexação, metadados, referencia, recuperação da informação, direitos autorais e preservação da informação digital; e) principais fontes de reuniões técnicas especificas, lista de discussão, grupos e centros de estudos, cursos e treinamento.

OGMA	OGMA WEB
1. bibliografia internacional seletiva e anotada sobre bibliotecas digitais 2. os seguintes aspectos 3. principais autores 4. o período de 1945-1985 5. a biblioteca do futuro 6. conceituação de biblioteca digital 7. projetos em andamento na alemanha	1. bibliografia internacional seletiva e anotada sobre bibliotecas digitais 2. os seguintes aspectos 3. principais autores 4. a biblioteca do futuro 5. o período de 1945-1985 6. conceituação de biblioteca digital 7. projetos em andamento na alemanha

8. austrália 9. brasil 10. canadá 11. dinamarca 12. espanha , estados unidos , franca 13. holanda 14. japão , nova 15. zelândia 16. reino unido 17. suécia e vaticano 18. aspectos técnicos relativos 19. a construção de uma biblioteca digital 20. arquitetura do sistema 21. marcação de textos 22. desenvolvimento de coleções 23. catalogação 24. classificação 25. indexação 26. metadados 27. recuperação da informação , direitos autorais e preservação da informação digital 28. conversão de dados e escaneamento 29. principais fontes de reuniões técnicas 30. grupos e centros de estudos 31. cursos e treinamento 32. lista de discussão	8. austrália 9. brasil 10. canadá 11. dinamarca 12. espanha , estados unidos , franca 13. holanda 14. japão , nova 15. zelândia 16. suécia e vaticano 17. reino unido 18. aspectos técnicos relativos 19. a construção de uma biblioteca digital 20. arquitetura do sistema 21. marcação de textos 22. desenvolvimento de coleções 23. catalogação 24. classificação 25. indexação 26. metadados 27. recuperação da informação , direitos autorais e preservação da informação digital 28. conversão de dados e escaneamento 29. principais fontes de reuniões técnicas 30. grupos e centros de estudos 31. cursos e treinamento 32. lista de discussão

RESUMO 3: *FAGUNDES, Maria Lúcia Figueiredo; PRADO, Gilberto dos Santos. Videoteca digital: a experience a da videotexa multimeios do IA/UNICAMP. Transformação, Campinas, v.11, n3, p. 293-299, set./dez. 1999.*

TEXTO: Apresenta a implantação de recursos multimídia e interface Web no banco de dados desenvolvido para a coleção de vídeos da Videoteca Multimeios, pertencente ao Departamento de Multimeios do Instituto de Artes da UNICAMP. Localiza a discussão conceitual no universe das bibliotecas digitais e propõe alterações na configuração atual de seu banco de dados.

OGMA	**OGMA WEB**
1. pertencente 2. o instituto de artes da unicamp 3. a o departamento de multimeios do instituto de artes da unicamp 4. a implantação de recursos multimídia e interface web no banco de dados desenvolvido para a coleção de vídeos da videoteca multimeios	1. pertencente 2. o instituto de artes da unicamp 3. a implantação de recursos multimídia e interface web no banco de dados desenvolvido para a coleção de vídeos da videoteca multimeios 4. a o departamento de multimeios do instituto de artes da unicamp

5. a discussão conceitual no universe das bibliotecas digitais 6. alterações na configuração atual de seu banco de dados	5. a discussão conceitual no universe das bibliotecas digitais 6. alterações na configuração atual de seu banco de dados

RESUMO 4: *FAGUNDES, Maria Lúcia Figueiredo; PRADO, Gilberto dos Santos. Videoteca digital: a experiencia da videotexa multimeios do IA/UNICAMP. Transformação, Campinas, v.11, n3, p. 293-299, set./dez. 1999.*

TEXTO: Este artigo aborda a necessidade de adoção de padrões de descrição de recursos de informação eletrônica, particularmente, no âmbito da Embrapa Informática Agropecuária. O rural Mídia foi desenvolvido utilizando o modelo Dublin Core (DC) para descrição de seu acervo acrescido de pequenas adaptações introduzidas diante da necessidade de adequar-se especificidades meramente institucionais. Este modelo de metadados baseado no Dublin Core, adaptado para o Banco de Imagem, possui características que endossam a sua adoção, como simplicidade na descriçao dos recursos, entendimento semântico universal (dos elementos), escopo internacional e extensibilidade (o que permite sua adaptação as necessidades adicionais de descrição).

OGMA	**OGMA WEB**
1. este artigo 2. a necessidade de adoção de padrões de descrição de recursos de informação eletrônica 3. particularmente 4. o âmbito da embrapa 5. informática agropecuária 6. o rural mídia 7. dublin 8. o modelo 9. descrição de seu acervo acrescido de pequenas adaptações introduzidas diante de a necessidade 10. a ele especificidades 11. meramente institucionais 12. a sua adoção 13. entendimento semântico universal 14. este modelo de metadados baseado no dublin 15. o banco de imagem 16. como simplicidade na descriçao dos recursos 17. dos elementos 18. escopo internacional e extensibilidade 19. sua adaptação 20. as necessidades adicionais de descrição	1. este artigo 2. a necessidade de adoção de padrões de descrição de recursos de informação eletrônica 3. particularmente 4. o âmbito da embrapa 5. informática agropecuária 6. o rural mídia 7. dublin 8. o modelo 9. descrição de seu acervo acrescido de pequenas adaptações introduzidas diante de a necessidade 10. a ele especificidades 11. meramente institucionais 12. a sua adoção 13. entendimento semântico universal 14. este modelo de metadados baseado no dublin 15. o banco de imagem 16. como simplicidade na descriçao dos recursos 17. dos elementos 18. escopo internacional e extensibilidade 19. sua adaptação 20. as necessidades adicionais de descrição

RESUMO 5: *CHATAIGNIER, Maria Cecilia Pragana; SILVA, Margareth Prevor, Biblioteca digital: a experiência do Impa. Ciência da Informação Brasília, v. 30, n. 3, p 7-12, set/dez.*

2001.

TEXTO: Relato da experiência do Impa na informatização de sua biblioteca, utilizando o software Horizon, e na construção de um servidor de preprints (dissertações de mestrado, teses de doutorado e artigos ainda não publicados) através da participação no projeto internacional Math-Net.

OGMA	**OGMA WEB**
1. a experiência do impa na informatização de sua biblioteca 2. o software 3. horizon 4. a construção de um servidor de preprints 5. relato da experiência do impa na informatização de sua biblioteca 6. dissertações de mestrado 7. teses 8. artigos aindanão publicados 9. internacional math-net 10. através da participação no projeto 11. a participação no projeto	1. a experiência do impa na informatização de sua biblioteca 2. o software 3. horizon 4. a construção de um servidor de preprints 5. relato da experiência do impa na informatização de sua biblioteca 6. dissertações de mestrado 7. teses 8. artigos aindanão publicados 9. ainda não publicados 10. internacional math-net 11. projeto internacional 12. através da participação no projeto 13. a participação no projeto

RESUMO 6: *MARCODENTES, Carlos Henrique; SAYAO, Luiz Fernando, Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. Ciência da Informação, Brasília, v.30, n.3, p24-33, set./dez.*

2001.

TEXTO: Descreve as opções tecnológicas e metodológicas para atingir a interoperabilidade no acesso a recursos informacionais eletrônicos, disponíveis na Internet, no âmbito do projeto da Biblioteca Digital Brasileira em Ciência e Tecnologia, desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia(IBCT). Destaca o impacto da Internet sobre as formas de publicação e comunicação em C&T e sobre os sistemas de informação e bibliotecas. São explicitados os objetivos do projeto da BDB de fomentar mecanismos de publicação pela comunidade brasileira de C&T, de textos completos diretamente na internet, sob a forma de teses, artigos de periódicos, trabalhos em congressos, literatura "cinzenta", ampliando sua visibilidade e acessibilidade nacional e

internacional, e também de possibilitar a interoperabilidade entre estes recursos informacionais brasileiros em C&T, heterogêneos e distribuidos, através de acesso unificado via um portal, sem necessidade de o usuário navegar e consultar cada recurso individualmente.

OGMA	**OGMA WEB**
1. as opções tecnológicas e metodológicas 2. a recursos 3. informacionais eletrônicos , disponíveis na internet 4. a biblioteca digital brasileira em ciência e tecnologia , desenvolvido 5. instituto brasileiro de informação em ciência e tecnologia 6. a interoperabilidade no acesso 7. o âmbito do projeto da biblioteca digital brasileira em ciência e tecnologia , desenvolvido 8. ibct 9. a internet 10. publicação e comunicação em c&t 11. os sistemas de informação e bibliotecas 12. o impacto da internet 13. a bdb 14. mecanismos de publicação por a comunidade brasileira de c&t 15. textos completos 16. a internet 17. artigos 18. trabalhos em congressos 19. literatura cinzenta 20. sua visibilidade e acessibilidade nacional 21. a interoperabilidade entre estes recursos 22. informacionais brasileiros em c&t , heterogêneos e distribuidos 23. necessidade 24. cada recurso 25. a forma de teses 26. através de acesso unificado 27. explicitados os objetivos do projeto da bdb	1. as opções tecnológicas e metodológicas 2. a recursos 3. informacionais eletrônicos , disponíveis na internet 4. a biblioteca digital brasileira em ciência e tecnologia , desenvolvido 5. instituto brasileiro de informação em ciência e tecnologia 6. a interoperabilidade no acesso 7. o âmbito do projeto da biblioteca digital brasileira em ciência e tecnologia , desenvolvido 8. ibct 9. a internet 10. publicação e comunicação em c&t 11. os sistemas de informação e bibliotecas 12. o impacto da internet 13. a bdb 14. mecanismos de publicação por a comunidade brasileira de c&t 15. textos completos 16. a internet 17. artigos 18. trabalhos em congressos 19. literatura cinzenta 20. sua visibilidade e acessibilidade nacional 21. a interoperabilidade entre estes recursos 22. informacionais brasileiros em c&t , heterogêneos e distribuidos 23. necessidade 24. cada recurso 25. a forma de teses 26. através de acesso unificado 27. explicitados os objetivos do projeto da bdb

III. Validação : Similaridade OGMA

RESUMO 1 e 2: MAIA, L. C.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência Informação, Belo Horizonte,** v. 15, p. 154-172, 2010.

TEXTO: Esta pesquisa verificou se ocorre aprimoramento na classificação de documentos eletrônicos com o uso de técnicas e algoritmos de mineração de texto (análise de texto) utilizando-se, além das palavras, sintagmas nominais como indexadores. Utilizaram-se duas ferramentas nos experimentos propostos desta pesquisa o OGMA e a WEKA. O OGMA foi desenvolvido pelos autores para automatizar a extração dos sintagmas nominas e o cálculo do peso de cada termo na indexação dos documentos para cada um dos seis métodos propostos. A WEKA foi utilizada para analisar os resultados encontrados pelo OGMA utilizando aos algoritmos de agrupamento e classificação, SimpleKMeans e NaiveBayes, respectivamente, obtendo um valor percentual indicando quantos documentos foram classificados corretamente. Os métodos com melhores resultados foram o de termos sem stopwords e o de sintagmas nominais classificados e pontuados como descritores.

RESUMO 3: SILVA, E. M. D.; SOUZA, R. R. Information retrieval system using Multiwords Expressions (MWE) as descriptors. **JISTEM - Journal of Information Systems and Technology Management,** v. 9, p. 213-234, 2012. ISSN 1807-1775. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-17752012000200003&nrm=iso17752012000200003&nrm=iso >.

TEXTO: This paper aims to propose an alternative method for retrieving documents using Multiwords Expressions (MWE) extracted from a document base to be used as descriptors in search of an Information Retrieval System (IRS). In this sense, unlike methods that consider the text as a set of words, bag of words, we propose a method that takes into account the characteristics of the physical structure of the document in the extraction process of MWE. From this set of terms comparing pre-processed using an exhaustive algorithmic technique proposed by the authors with the results obtained for thirteen different measures of association statistics generated by the software Ngram Statistics Package (NSP). To perform this experiment was set up with a corpus of documents in digital format.

RESUMO 4: SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. **Perspectivas em Ciência da Informação,** v. 10, n. 2, 2008. ISSN 1981-5344.

TEXTO: Desde que se tornaram inviáveis em alguns contextos os processos manuais de indexação de documentos, buscam-se alternativas eficazes que possibilitem a representação automática dos assuntos principais desses documentos. Os processos mais comuns de indexação automática descrevem os documentos através de uma lógica simplista advinda da análise de freqüência das palavras que neles ocorrem. Buscando propor processo de indexação mais eficaz, que analise as palavras e expressões no âmbito de seus contextos

lingüísticos, três pressupostos são definidos: a) a utilização de sintagmas nominais como descritores apresenta vantagens em relação ao uso de palavras- chave; b) a extração de sintagmas nominais de textos de documentos digitalizados é possível e viável com ferramentas tecnológicas atualmente disponíveis e c) é possível estabelecer processo automatizado e eficaz para escolha de descritores significativos para documentos digitalizados, utilizando sintagmas nominais. O objetivo da pesquisa é apresentar uma metodologia para viabilizar o processo de atribuição de descritores a textos digitalizados – indexação – através da extração de sintagmas nominais e da análise de fatores como a freqüência de ocorrência desses sintagmas nominais nos textos dos documentos, no conjunto dos documentos; a estrutura dos sintagmas nominais; o nível dos sintagmas nominais e a ocorrência desses em tesauro de um campo de conhecimento específico. Para atingir esse objetivo são analisados (a) um corpus de 15 documentos dos quais foram extraídos os sintagmas nominais manualmente, para testar o processo de extração automática e (b) um corpus de 60 documentos provenientes de publicações eletrônicas da área de ciência da informação. A metodologia proposta foi aplicada inicialmente a parte do corpus para validação e parametrização das variáveis do algoritmo, e, então, novamente aplicada, com alterações, à totalidade do corpus. Os resultados apresentados demonstraram grande pertinência dos descritores atribuídos aos documentos e permitiram concluir que a metodologia obtém sucesso inequívoco nas condições estudadas.

IV. Algoritmo de Agrupamento

```
<?php
     //Rodrigo - 20/08/2013
     //Implementado método de similaridade múltipla de documentos
public function SimilaridadeMultipla($arquivos,$fout){
        //Monta a matriz dimensional de distâncias
        $fp = fopen($fout,"w+");
        $cabecalho=';';
        $conteudo='';
        foreach($arquivos as $k1=>$a1){
$cabecalho.=$a1["doc"].';';
            $conteudo.=$a1["doc"].';';
            $temp=array();
            if(! is_array($a1["termos"]) || count($a1["termos"])==0){
echo "Extraia os termos do arquivo ".$a1["doc"]." antes de realizar o
cálculo de similaridade.";                 return false;
            }
            foreach($arquivos as $k2=>$a2){
if($k1!=$k2){
                    if(! is_array($a2["termos"]) ||
count($a2["termos"])==0){
                        echo "Extraia os termos do arquivo
".$a2["doc"]." antes de realizar o cálculo de similaridade.";
return false;
                    }
                    $temp[$a2["doc"]] = abs(1 - $this-
>Similaridade($a1["termos"],$a2["termos"],$a1["doc"],$a2["doc"],""))
;
                    $conteudo.=$temp[$a2["doc"]] .';';
                }else{
                    $conteudo.='0;';
                }
            }
            $conteudo.="\r\n";
            $a1["sim"]=$temp;
        }
        fwrite($fp,$cabecalho."\r\n".$conteudo);
fclose($fp);
        return $this->Agrupamento($arquivos);
    }
    public function Agrupamento($arquivos){
          $grupo=1;
          while(!$this-
>concluiuAgrupamento($arquivos)){
               //Identifica os pares de acesso
               $semGrupo=0;
               foreach($arquivos as $k=>$a){
                    if((!isset($a["grupo"]) || $a["grupo"]=="")){
                         $semGrupo++;
                         $arq = $k;
                         $menor=1;
                         $par="";

                         foreach($a["sim"] as $doc => $distancia){
if($distancia<$menor){
$menor=$distancia;
```

```
                                    $par=$doc;
                                }
                            }
                            $a["par"]=$par;
                        }
                    }
if($semGrupo>1){
            //Acha Alpha e o par de início
            $alpha = 0;
            $menor = 1;
            $par00 = "";
$par01 = "";
            foreach($arquivos as $k=>$a){
                if((!isset($a["grupo"]) || $a["grupo"]=="") &&
$a["sim"][$a["par"]] > $alpha){
                    $alpha = $a["sim"][$a["par"]];
                }
                if((!isset($a["grupo"]) || $a["grupo"]=="") &&
$a["sim"][$a["par"]] < $menor){
                    $par00 = $a["doc"];
                    $par01 = $a["par"];
                    $menor = $a["sim"][$a["par"]];
                }
            }
            //Coloca os 2 arquivos no grupo
foreach($arquivos as $k=>$a){
                if($a["doc"]==$par00 ||  $a["doc"]==$par01){
                    $a["grupo"]=$grupo;
                }
            }
            //Verifica aceitação dos elementos no grupo
            $semGrupo=0;
            foreach($arquivos as $k=>$a){
                    if((!isset($a["grupo"]) || $a["grupo"]=="") &&
$a["doc"]!=$par00 && $a["doc"]!=$par01){
                        $distancia = (($this-
>obtemDistancia($arquivos,$par00,$a["doc"]) +
$this>obtemDistancia($arquivos,$par01,$a["doc"])) - $menor);
if($distancia <= $alpha){
                        $a["grupo"]=$grupo;
                    }
}
                if((!isset($a["grupo"]) || $a["grupo"]=="")){
                    $semGrupo++;
                    $arq = $k;
                }
            }
        }
        //Atribui o grupo mais proximo ao último elemento não
agrupado          //(no caso do número de documentos submetidos ser
impar)         if($semGrupo==1){
            foreach($arquivos as
$k=>$a){                 if($k==$arq){
$menor=1;
                    foreach($a["sim"] as
$doc=>$distancia){
```

```
if($distancia<$menor){
$a["grupo"] = $this-
>obtemGrupo($arquivos,$doc);
                            $menor = $distancia;
                        }
                    }
                }
            }
        }
        $grupo++;

}
        return $arquivos;
    }      private function
imprimeGrupo($arquivos){
foreach($arquivos as $k=>$a){
            echo $a["doc"]." - ".$a["grupo"]."<br/>\r\n";
        }
}
    private function
obtemGrupo($arquivos,$doc){
foreach($arquivos as $k=>$a){
if($a["doc"]==$doc){                    return
$a["grupo"];
            }
        }
}
    private function
obtemDistancia($arquivos,$docA,$docB){
foreach($arquivos as $k=>$a){
if($a["doc"]==$docA){                    return
$a["sim"][$docB];
            }
        }
}
    private function concluiuAgrupamento($arquivos){
foreach($arquivos as $k=>$a){
            if(!isset($a["grupo"]) || $a["grupo"]==""){
return false;
            }
}         return
true;
    }

?>
```